STUDIES ON GENOMIC G-QUADRUPLEXES

by

Julian Huppert

Trinity College, Cambridge

A dissertation submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy at the
University of Cambridge

December 2004

PREFACE

The work described in this thesis was carried out by the author in the Department of Chemistry, University of Cambridge, between October 2001 and December 2004 under the supervision of Dr Shankar Balasubramanian and supported by a BBSRC studentship and the Isaac Newton Trust. It is original except as indicated by reference and has not been submitted for any other degree at this or any other university. The thesis does not exceed 60 000 words in length.

Julian Huppert

Trinity College, Cambridge
December 2004

ACKNOWLEDGEMENTS

My thanks must go particularly to Dr Shankar Balasubramanian for his ongoing support and supervision throughout this project. Thanks especially for allowing me the freedom to explore and develop this field in whichever form seemed to be suitable – and for the necessary patience! I am also indebted to him for persuading the Trinity Fellowship to award me a Junior Research Fellowship to allow me to continue this work.

The Balasubramanian Group, current and past, has been friendly and supportive, and has made labwork bearable, if not enjoyable. My particular thanks for friendship and discussions go to Dr Yamuna Krishnan-Ghosh, Dr Jeremy Green, Dr Sylvain Ladame, Jen Hake and Shuizi 'Rachel' Yu. Thanks for technical assistance to Dr Sylvain Ladame for performing the SPR experiments, Dr James Redman for performing the ELISA experiments, and Jen Hake, for collaborating on the fluorescence experiments. Rachel has been a wonderful Part III student, and a pleasure to supervise. My excellent team of proofreaders, Shankar, Sylvain, Yamuna, Jeremy, James and Caroline, deserve much praise and credit – though none of the blame!

This project has been heavily interdisciplinary, and I would like to thank many people outside the group for their assistance. Pascale Hazel performed the molecular modelling studies in Chapter 2, and has been very helpful in discussions and to bounce ideas off. Simon Rodgers taught me how to program, and was always available to explain why the code I had just written wasn't working – thanks! Thanks also to the Sarahs, both Teichmann and Kummerfeld, for discussions about cross-genome quadruplexes. Manolis Dermitzakis and Richard Durbin at the Sanger Institute have been extremely stimulating, and I look forward to working further with them. Dr Marta Zlatic has taught me everything I know about fruit flies, and has collected and fed many on my behalf. Profs Geoff Grimmett and Herbert Huppert assisted with the mathematical analysis in appendix A. Dr Patsy Altham assisted with the statistical approaches to loop lengths, and generated the mosaic plot used therein. Prof Felicia Huppert provided me with an extra reason to celebrate completing this thesis.

Almost my entire PhD, I have been fortunate to have Caroline Wright to support me, to be with and to love. She is a highly multitalented person, who seems to be prepared to put up with a vast amount – including occasional politics! I'm sure she will succeed in whatever she does next.

However, there are people who have done even more, who have loved, cherished, supported and encourage me since I was born – my family. Mum, Dad and Row: Thank you all very much.

SUMMARY

The double-helical structure of DNA is well known, and the basis of modern genetics and molecular biology. However, DNA is polymorphic, and can adopt a variety of different structures. In this dissertation, I focus on a four-stranded structure that can be formed by guanine-rich sequences. These are known as G-quadruplexes or G-tetrads.

I begin by investigating the structural properties of this structure, and how sensitive it is to mutations and deletions. I then consider the structural properties of the loops that join the four strands, and develop an understanding of how the length of the loops affects the stability of the structures, and also which folding pattern they adopt.

Using the above results and some other considerations, I then develop a 'folding' rule, which predicts which sequences are expected to form quadruplexes under physiological conditions.

Using this rule, I identify a number of putative quadruplex sequences in the promoter regions of a selection of oncogenes and develop a model for how these structures could be exploited as a drug target for gene regulation. The identified structures are characterised biophysically, and drug binding *in vitro* is demonstrated.

An *in vivo* system using the fruit fly *Drosophila Melanogaster* is used to test whether drugs can be used to target a quadruplex in the promoter region of a key neuronal gene.

I then address the hypothesis that quadruplexes could be a natural mechanism for gene regulation (or other functions). In order to investigate this, I develop a technique to search rapidly the entire human genome for quadruplex-forming sequences using the folding rule derived above. This identifies 350,000 potential

sequences in the human genome. This is compared to the number expected if the DNA sequence was purely random (solved analytically), and using a simple Markov model for the human genome, showing that there are fewer such sequences than expected.

Statistical study of the correlations between the lengths of the three loops formed by potential genomic quadruplexes show strong correlations, and these may be explained in terms of the folding pattern of the quadruplexes. This provides the first evidence of wide-scale presence of actual quadruplexes in the genome, and allows the calculation of a lower estimate for the number present.

Co-location of Single Nucleotide Polymorphisms (SNPs) and quadruplex sequences is studied, with especial focus on those that have been correlated with diseases. A number of interesting clinical observations that could be attributed to quadruplex formation are investigated, including the COL1A1 osteoporosis gene.

A number of quadruplex sequences are found to be conserved between human and mice, in similar positions. This is investigated further, and used to provide further evidence that there is some significance to these sequences. The role of transcription factors in binding these quadruplexes is discussed.

In summary, this thesis broadens the quadruplex DNA field to consider their prevalence throughout the genome, develops a number of potential drug candidates, and demonstrates how important some of these sequences may be for biological function.

STUDIES ON GENOMIC G-QUADRUPLEXES

CONTENTS

ABBREVIATIONS

A	adenine
A_x	absorbance at x nm
ANOVA	analysis of variance
bp	base-pairs
C	cytosine
CD	circular dichroism
DMS	dimethylsulphate
DNA	deoxyribonucleic acid
dsDNA	double-stranded DNA
d(xxx)	DNA sequence xxx
EC_{50}	Concentration of ligand require to achieve 50% inhibiton
ELISA	enzyme-linked immunoabsorbant assay
FRET	fluorescence resonance energy transfer
G	guanine
ΔG	change in Gibb's free energy
GFP	green fluorescent protein
ΔH	change in enthalpy
K_d	binding constant
LogP	hydrophobicity measure
LNA	locked nucleic acid
MD	molecular dynamics
MW	molecular weight
natoms	number of atoms
NHE	nuclease hypersensitivity element
NMR	nuclear magnetic resonance
nrotb	number of rotatable bonds
nOHNH	number of hydrogen bond donors
nON	number of hydrogen bond acceptors

nviolations	number of violations of the Lipinski rules
-p-	phosphate
PAGE	polyacrylamide gel electrophoresis
PDB	protein data bank
PNA	peptide nucleic acid
PQS	putative quadruplex sequence
PSA	polar surface area
rms	root mean square
RNA	ribonucleic acid
ΔS	change in entropy
SNP	single nucleotide polymorphism
SPR	surface plasmon resonance
ssDNA	single-stranded DNA
T	thymine
TBA	thrombin binding aptamer
TF	Transcription factor
T_m	melting temperature
U	uridine
UTR	Untranslated region
UV	ultraviolet

TABLE OF FIGURES AND TABLES

Figures

Tables

CHAPTER 1

Introduction

1.1 DNA structures

DNA is a polymer comprising a string of one of four deoxynucleotides, each consisting of a pentose sugar with a phosphate group on the 5' hydroxyl, and a heterocyclic base attached to the 1' position (see figure 1.1.1). The base may be adenine (A), cytosine (C), guanine (G) or thymine (T). Watson and Crick realised[1] that these nucleobases were capable of forming specific base-pairs, A with T, and G with C, held together by specific hydrogen bonding patterns (see figure 1.1.2). This observation explained the Chargaff rules,[2] in which Chargaff observed that there were equal compositions of A and T, and of G and C in genomic DNA.

The base-pairing rules, together with the X-ray diffraction results of Franklin and Wilkins,[3,4] led them to deduce that the normal structure of genomic DNA is a double helix (see figure 1.1.3). This understanding has formed the basis for the modern-day understanding of genetics and molecular biology,[5] as, in the immortal words of Watson and Crick it 'immediately suggests a possible copying mechanism for the genetic material'.[1]

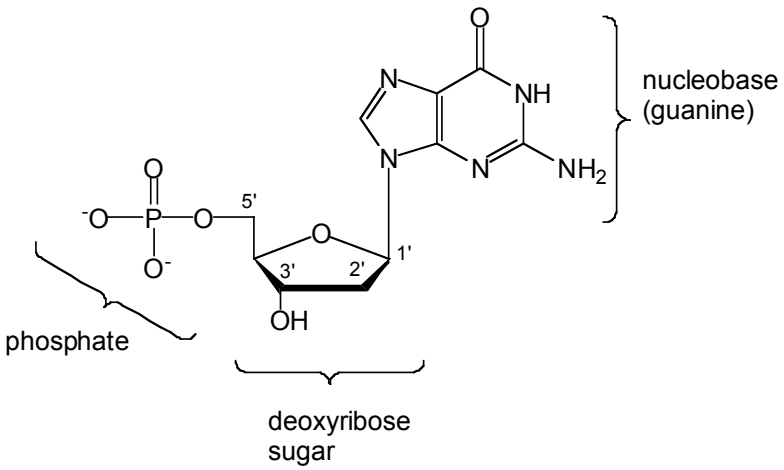


Figure 1.1.1. Structure of a DNA nucleotide. It is comprised of a 2'-deoxyribose sugar with a phosphate attached to the 5' hydroxyl, and one of the four nucleobases attached to the 1' position. The 3' hydroxyl can be used to attach the phosphate of the next nucleotide. Ribonucleic acid (RNA) is similar in structure, except it also has a 2' hydroxyl group and one different nucleobase – uracil in place of thymine.

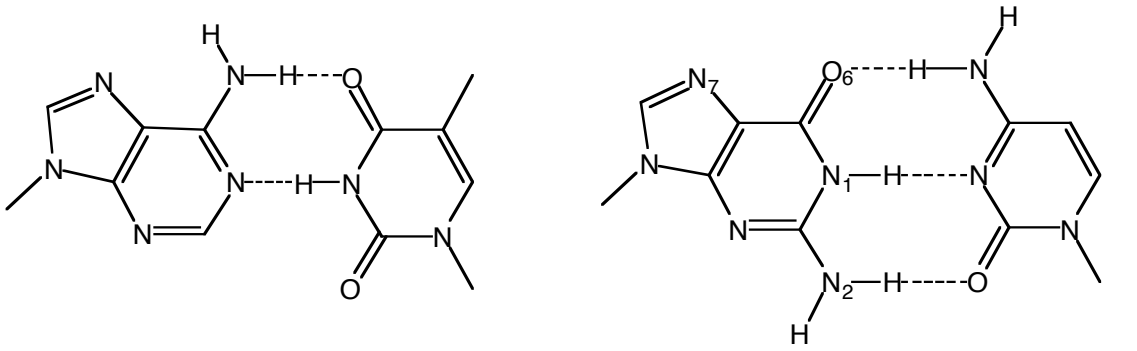


Figure 1.1.2: Watson-Crick base-pairs formed between A and T (left, two hydrogen bonds) and between G and C (right, three hydrogen bonds). Significant positions in guanine are labelled for future reference.

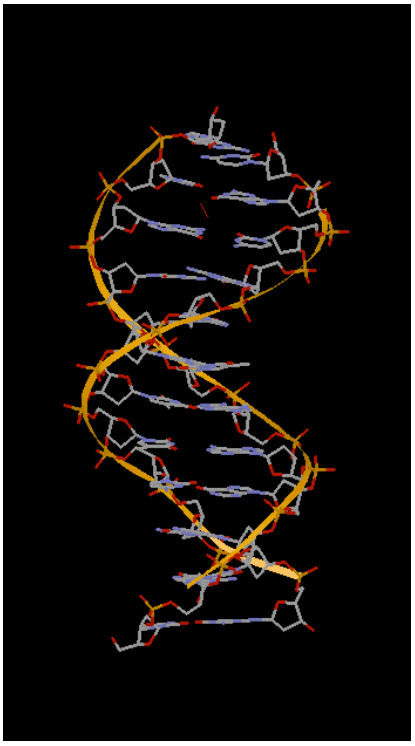


Figure 1.1.3. Structure of duplex B-DNA. The phosphate backbone is highlighted in gold, and base pairing can be observed. PDB entry 1BNA. Drawn using RasMol.[6]

1.1.1 Guanosine can aggregate

Studies on guanosine nucleotides in the early 1900's showed that guanosine mononucleotides formed viscous gels at millimolar concentrations in water, in contrast to the other nucleotides, which showed no such behaviour.[7] In 1962, X-ray diffraction studies[8] suggested that the guanosine was forming a tetrameric structure with a square co-planar array of four guanines, using the N_1, N_7, O_6 and

N_2 of each guanine base, with a metal ion (or occasionally other monovalent cation) in the centre (figure 1.1.4). This hypothesis was later confirmed by other biophysical studies.[9,10].

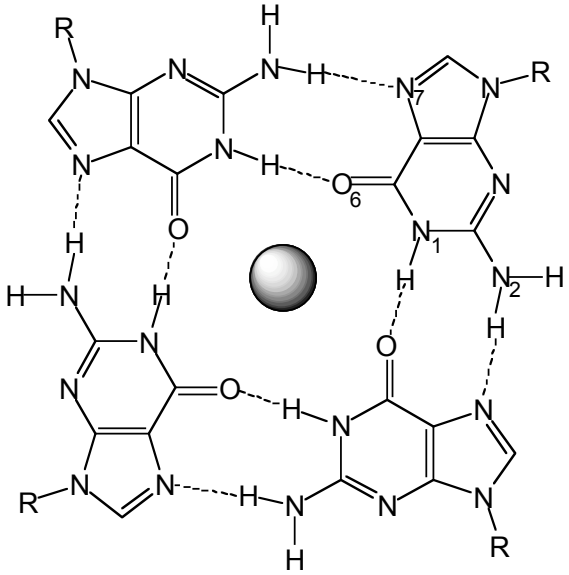


Figure 1.1.4. Structure of a guanine tetrad. Hydrogen bonding occurs using both the Watson-Crick and Hoogsteen faces, resulting in a circular planar network of hydrogen bonds, averaging out at two per guanine (cf 1.5 per guanine in duplex DNA). The metal ion is bound by the surrounding oxygen lone pairs. Key positions are labelled for one guanine (top right).

1.1.2 G-quadruplexes are formed from stacked tetramers

In DNA (and RNA), consecutive G-tetrads can stack to form higher-order structures,[11-15] held together by π-π stacking and in most cases internal monovalent cations. These structures have become known as G-quadruplexes, G-quartets, G-tetraplexes and G4-DNA, but in this thesis I shall use the terminology G-quadruplex. They are helical in nature, with a right handed twist of approximately 12 bases per turn,[16,17] compared to 10 bases per turn for B-DNA.[5] RNA sequences can also form quadruplexes,[18] as can the unnatural DNA analogues protein nucleic acid (PNA)[19,20] and locked nucleic acid (LNA).[21] A schematic of a quadruplex is shown in figure 1.1.5, and views of a quadruplex are shown in figure 1.1.6.

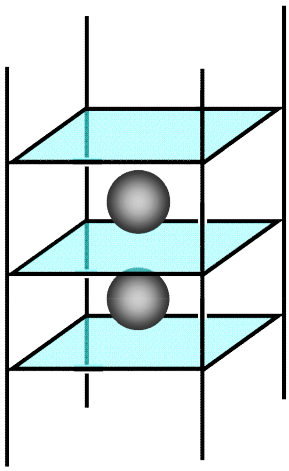
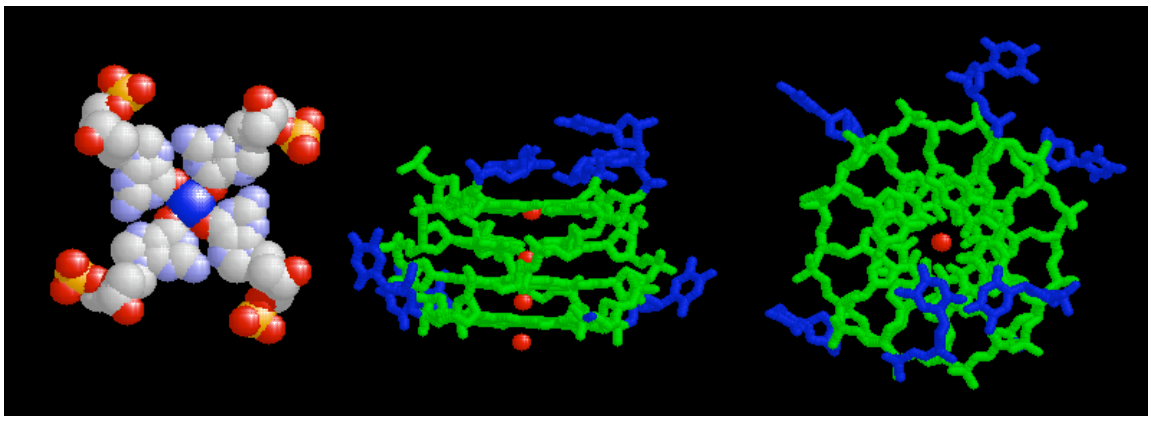


Figure 1.1.5. Schematic structure of a G-quadruplex with intercalated metal ions. The helicity has not been represented in this diagram for clarity. Cyan squares represent the tetrads.



a) b) c)

Figure 1.1.6 Crystal structure of a quadruplex. From left to right: a) spacefilling representation of a G-tetrad, including the intercalated Na^+ (blue). b) side view of a parallel four stranded quadruplex. c) top view of the same quadruplex. Guanines are shown in green, thymines in blue and Na^+ in red. Data are taken from PDB entry 244D (ref 22). The sequence is d(TGGGGT). The crystal unit contains four such quadruplexes, only one of which is shown. The bottom Na^+ in b) is shared between two of the four quadruplexes.

1.1.3 Quadruplex structures

G-quadruplexes can adopt a large number of different polymorphic structures.[12,23] They are dynamic structures, and in many cases can form alternative conformations under different conditions[24] or even exist as an equilibrium mixture.[25] All quadruplexes contain a core set of stacked tetrads, which defines them as a class, but otherwise can be very variable. The different

structures may be classified principally according to their stoichiometry and strand polarity as well as their topological form.

1.1.3.1 Stoichiometry and Strand polarity

G-quadruplexes can be formed by the association of one,[26-29] two[29,30] or four[22,31,32] strands, forming either inter- or intramolecular species. It is in principle possible to form 3-stranded species, with a hairpin and two single strands, but to date these have not been reported and are unlikely to be stable with respect to bi- and tetra-molecular species. Which of these structures is formed depends on the sequence of the strand,[33] on the strand concentration[34] and on the cations present.[35]

Within each of these structures, it is possible to have a number of combinations of strand orientations (3'-5' direction): all parallel,[29] three parallel and one anti-parallel,[26,36] two antiparallel pairs of adjacent parallel strands,[27,37] or alternating anti-parallel strands.[28] All of these have been observed experimentally and characterised by NMR or X-ray crystallography and are illustrated in figure 1.1.7.

1.1.3.2 Position of loops

For dimeric and intramolecular G-quadruplexes there is the additional issue of the geometry of the loops connecting the strands. Each connecting loop may connect parallel or antiparallel strands, either to adjacent strands, diagonally opposite strands, or to the other end of the structure, via a double chain reversal. A few sample structures are shown below in figure 1.1.7, all of which have been obtained experimentally. The length and sequence of the loops is also highly variable; loops commonly consist of 2–6 bases,[38] and for telomeric sequences (see next section) are rich in thymine.[17,39]

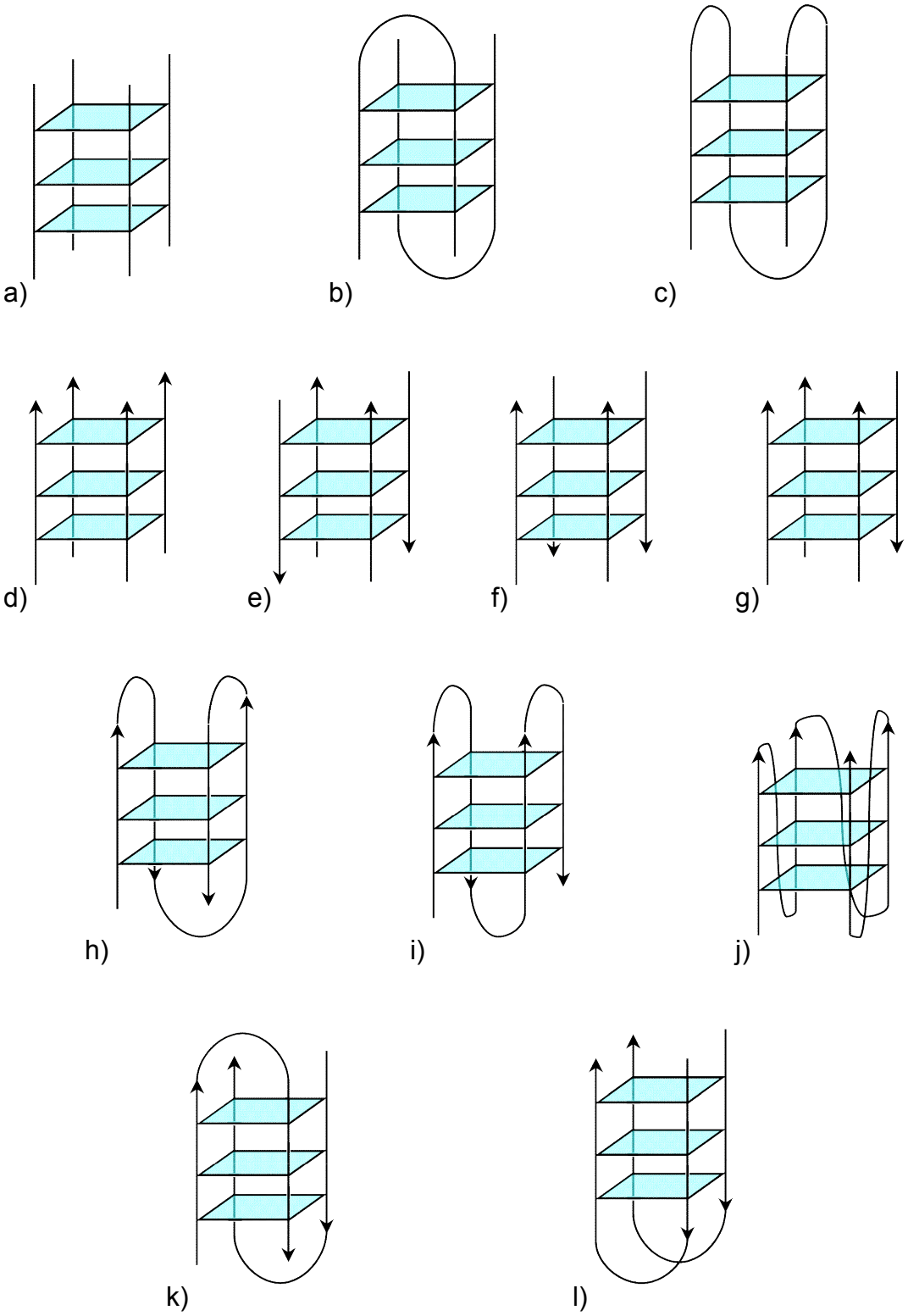


Figure 1.1.7. Various possible topologies for quadruplexes. Top row: stoichiometries: a) tetramer, b) dimer, c) intramolecular. Second row: strand orientations: d) all parallel, e) alternating antiparallel, f) adjacent parallel, g) three parallel, one antiparallel. Third row: sample of experimentally determined intramolecular forms: h) antiparallel with lateral loops, i) antiparallel with diagonal loop, j) parallel with double chain reversal loops. Fourth row: sample of experimentally determined dimeric (hairpin) forms: k) loops on opposing faces, l) loops on the same side.

There is one reported example of a double chain reversal loop containing no bases (as well as other unusual features).[40] Quadruplexes with longer loops have also been observed,[41,42] although one of these was made from an artificial sequence with extensive base pairing within extended loops.[42]

1.2 Telomeres and quadruplexes

Eukaryotic cells have linear chromosomes, and therefore need a mechanism to differentiate chromosome ends from double-strand breaks.[43] Also, the ends of a DNA strand cannot be replicated, known as the 'end-replication problem', and arises because of the mechanics of the replication process. DNA is made as a series of Okazaki fragments in the 5'-3' direction, which are then ligated together. However, the very last portion cannot be replicated, because that position is occupied by an RNA primer which is then deleted (see figure 1.2.1).[5] Consequently, the chromosomes shorten after each round of replication.

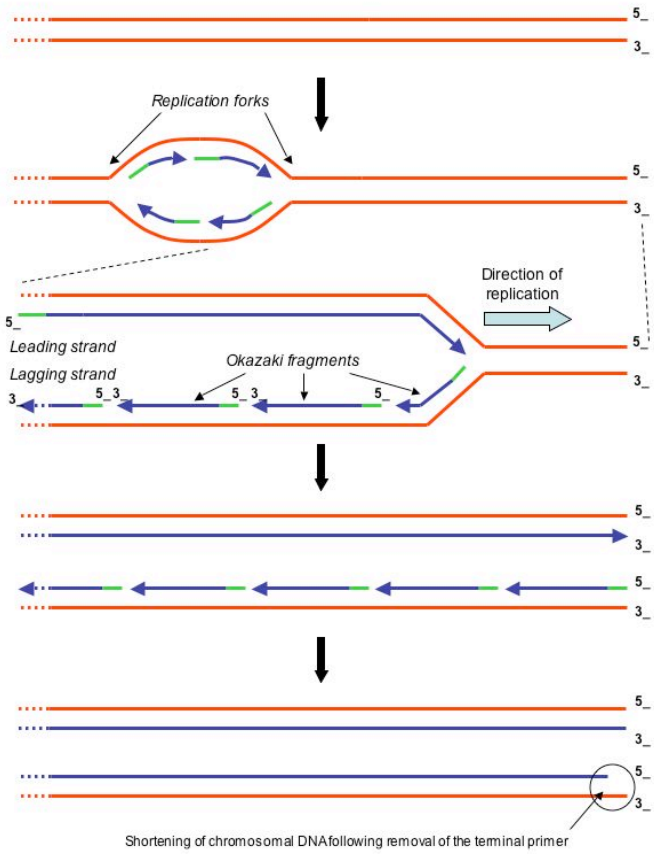


Figure 1.2.1. The end replication problem. The replication bubble migrates to the end of the chromosome, with continuous replication in the 5'-3' direction on one strand, and discontinuous replication via a series of Okazaki fragments on the other. The RNA primers initiating the Okazaki fragments are digested and the fragments ligated, but this leaves a gap at the 5' end of the new strands.

The solution to this problem in almost all eukaryotes is to have a long stretch of DNA containing repeats,[44] to which bind a number of specific capping proteins,[45-47] in order to mark it out as an end point rather than a breakage site. These structures at the chromosome ends are called telomeres. In order to lengthen this sequence after replication has shortened it, they use a reverse transcriptase, telomerase,[48,49] to replicate this sequence using an internal RNA template.

The sequence of telomeric repeats are in all cases found to be highly G-rich.[39] The repeat sequence varies only slightly among eukaryotes, being d(T_2G_4) for *Tetrahymena*, d(T_4G_4) for *Oxytricha*, and d(T_2AG_3) for all vertebrates. In human somatic cells, these repeats continue for around 5-8 kilobases, with a single-stranded 3' overhang of 100-200 bases (figure 1.2.2).[50] All of these sequences are capable of forming G-quadruplexes *in vitro*, and it is proposed that the single-stranded overhang exists in this form *in vivo*.[43,51] This has only been demonstrated conclusively *in vivo* in the case of *Stylonychia lemnae* telomeres, which were specifically bound by an antibody raised against an antiparallel quadruplex.[52]

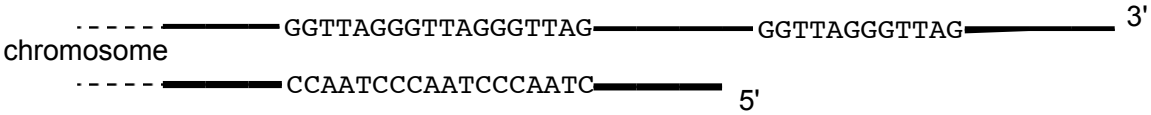


Figure 1.2.2. Schematic of a vertebrate telomere. Solid lines represent an undefined number of copies of the repeated units. For a human somatic cell, the double-stranded length would be around 8 kilobases and the single-stranded length around 100-200 bases.

It is interesting to note that the complementary strand, which has the sequence d(C_3TA_2), can form a four-stranded *i*-motif structure, which is known to be stable at relatively low pH, and involved bonding between hemiprotonated cytosine pairs, [53] This raises the possibility that the duplex region could form quadruplexes with the *i*-motif formation stabilising the other strand. This competition between duplex and quadruplex structures has been explored,[54-57] and it has been shown that the quadruplex structures are favoured at low pH, but that the duplex is favoured at more physiological conditions. The energy difference between the

two forms, however, is not so substantial that a protein interaction could not stabilise either form under physiological conditions. The equilibrium between these forms can also be affected by molecular crowding, which promotes quadruplex formation[58] and the binding of small molecules capable of stabilising the quadruplex form more than the duplex form.[59]

Oligonucleotides containing telomeric repeat sequences have been extensively studied. It has been shown that when stabilised by small molecules[60] or by increased potassium ion concentrations[61] they inhibit the activity of telomerase (more accurately, they prevent the substrate from being a target for telomerase activity). A correlation between quadruplex stability and telomerase inhibition has been demonstrated for one series of ligands.[62] Replacement of some guanines with 7-deazaguanine, in which the N_7 of guanine, involved in quadruplex formation, has been replaced with a CH group, prevented quadruplex formation and restored telomerase activity.[61]

The *Oxytricha* telomeric sequence d(T_4G_4) has been shown to form an antiparallel intramolecular quadruplex,[26] or at sufficiently high concentrations a parallel tetramer.[37,63] The *Tetrahymena* telomeric sequence d(T_2G_4) has been shown to form both parallel and antiparallel quadruplexes.[64] The human (and vertebrate) telomeric sequence d(T_2AG_3) has been the subject of considerable interest.

In 1993, Wang and Patel[27] solved the structure of the human telomeric sequence d(AGGG(TTAGGG)$_3$) in the presence of Na^+ using NMR spectroscopy. They obtained a structure with two parallel and two antiparallel strands, connected by two lateral loops at one end and a diagonal loop at the other. (Figure 1.2.3)

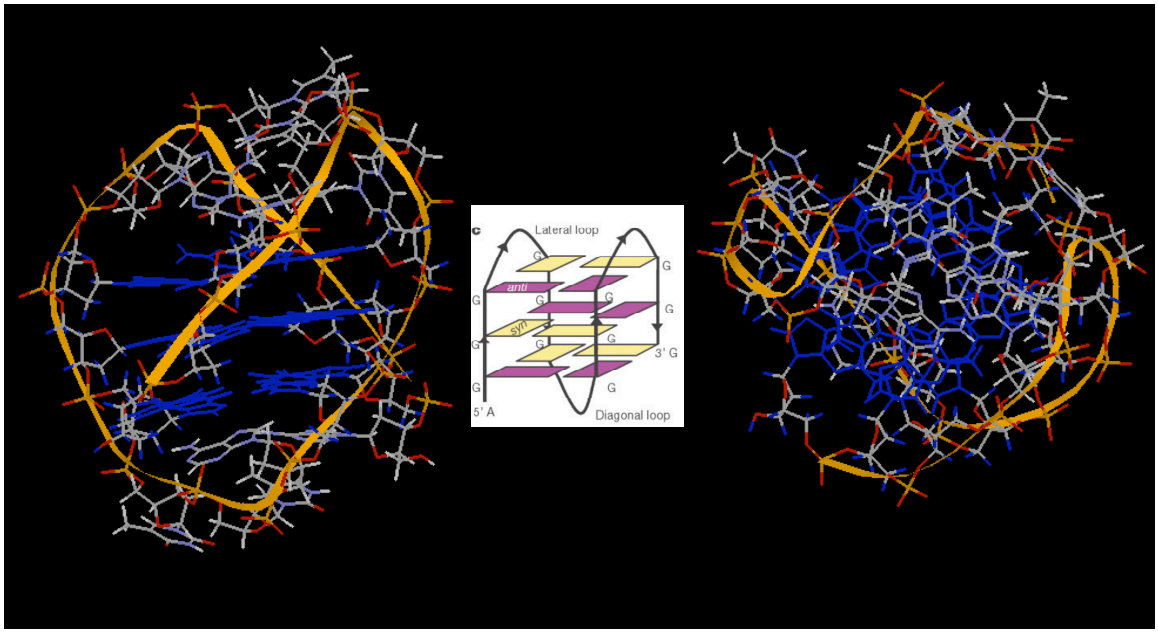


Figure 1.2.3. Wang-Patel antiparallel structure for the human telomeric quadruplex.[27] Left: side view. Right: top view. Center: schematic of folding pattern. Guanines are coloured blue and the phosphate backbone highlighted in gold. PDB entry 143D.

This structure has very accessible grooves for binding small molecules, as in duplex DNA, but binding to the top and bottom faces would clearly be difficult as the loops are covering the faces.

In 2002, Neidle and coworkers[29] published an X-ray crystal structure of the same sequence, but in the presence of K^+. In contrast to Wang and Patel's result,[27] they found a parallel-stranded structure, with double strand reversal loops bridging the tetrads in a way reminiscent of propeller blades. (Figure 1.2.4)

This structure has a very different morphology from the Wang-Patel structure, being much more disc-shaped than the other globular structure. The structure is also more symmetric. The π-rich tetrads at the ends are much more exposed for binding interactions, and the loops are also possible targets for molecular recognition.

The reasons for the discrepancy between the two results are not entirely clear. One explanation is that the different cations cause two different species to form, as has been seen for other sequences.[24,65] Another possibility is that the parallel structure is an artefact of the crystallisation process, as only one form will pack

into the crystal structure and hence be observed. Additionally, the Neidle study used a trisubstituted acridine to seed the quadruplex and induce crystallisation, which may explain the alternative form observed. However, the acridine is not observed in the final structure.

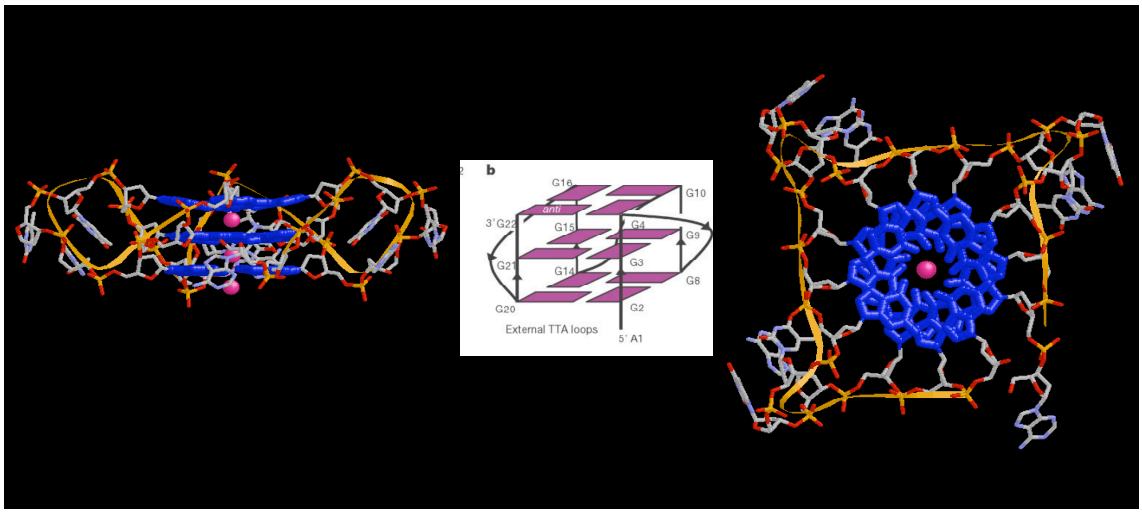


Figure 1.2.4. Neidle parallel structure for the human telomeric quadruplex.[29] Left: side view. Right: top view. Center: schematic of folding pattern. Guanines are coloured blue, potassium ions pink and the phosphate backbone highlighted in gold. PDB entry 1KF1.

In any event, it is clear that the two structures are both stable under relatively standard conditions, and there is now strong evidence that the two structures exist in equilibrium. Phan and Patel observed interconverting parallel and antiparallel species using NMR,[25] and it was shown by others in our group using FRET that two distinct forms existed in solution and could interconvert.[66]

1.3 Quadruplex binding molecules

There is considerable interest in identifying compounds that will bind and stabilise quadruplexes.[67,68] This is partly because there is pharmaceutical interest, as quadruplex binders have been shown to 'inhibit' the enzyme telomerase,[60] which has been shown to be expressed abnormally in 90% of cancer cells.[69] There are also a number of natural proteins that interact specifically with quadruplexes, and there is interest in understanding why these exist and what their functional roles are.

One particularly interesting method for observing quadruplex binding is the polymerase stop assay.[70] This technique relies on the observation that quadruplex sequences act as a steric block to strand replication; the amount of prematurely terminated sequence can be used as a measure of the effectiveness with which an applied drug binds to and stabilises the test quadruplex.

1.3.1 Small molecule binders

Among the ever-increasing list of small molecule quadruplex binders found to date are porphyrins,[71,72] ethidium derivatives,[73] N,N'-bis[2-(1-piperidino)ethyl)-3,4,9,10-perylenetetracarboxylic diimide (PIPER),[74-76] 3,3'-diethyloxadicarbocyanine (DODC),[77] acridine/acridone derivatives,[78] anthraquinones,[79] telomestatin[80] and peptide-hemicyanine conjugates[81] (Reviewed by Mergny [60]).

Many of these, of which a selection is shown in figure 1.4.1, have large π-surfaces, and it is believed they stack on top of the G-tetrads. Others, such as DODC[77] and perhaps the peptide-hemicyanine conjugates,[81] may bind in the grooves of the quadruplex. An extreme example is triethylene tetraamine (TETA), a linear polycation at neutral pH, which is believed to bind in the core of the quadruplex, displacing the metal ions that are normally bound.[82]

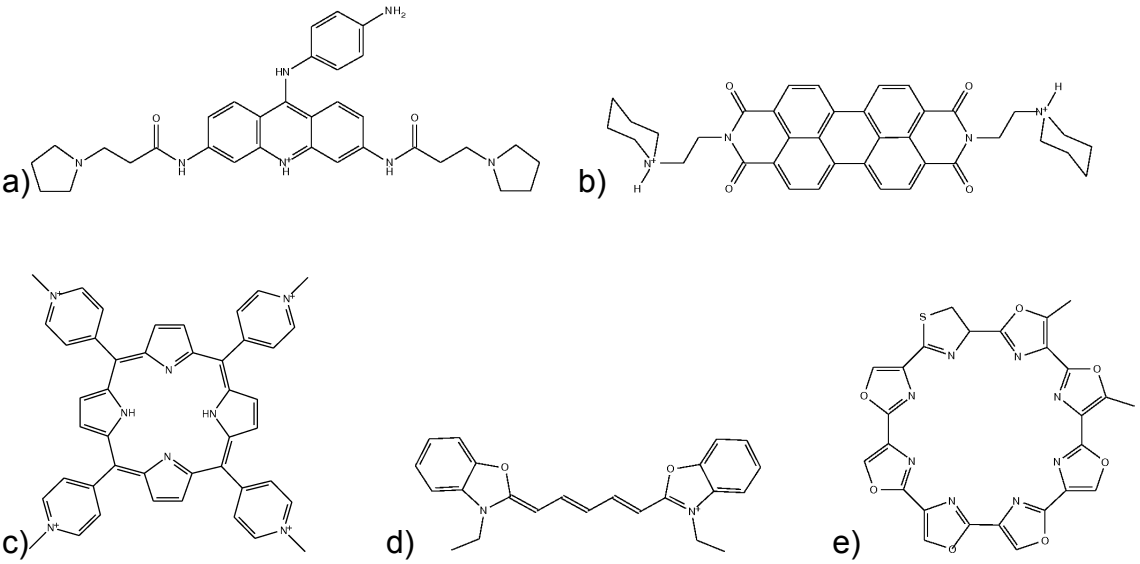


Figure 1.3.1. A selection of small molecule quadruplex binders. a)BRACO-19, a trisubstituted acridone. K_d 60 nM for quadruplex, EC_{50} 60 nM for telomerase.[78] b) PIPER. Inhibits telomerase above 20 μM.[74] c) TMPyP4, a tetrasubstituted porphyrin. K_d 36 μM for human telomeric repeat, 37 nM for *Oxytricha*.[71,72] d) DODC. K_d 9 μM against Oxytricha telomeric repeat. [77] e) Telomestatin. EC_{50} against telomerase 5 nM.[80]

One of the most interesting ligands is telomestatin, an extremely potent natural product that 'inhibits' telomerase.[80] This oligopeptide compound, extracted from *Streptomyces anulatus*, appears to stack well on top of the tetrads, and exhibits tight binding, as judged by the EC_{50} (concentration required to achieve 50% inhibition) for telomerase of 5 nM.[80] Its modular structure lends itself very well to systematic variation to optimise the binding and selectivity of this molecule.

Two structures have been published containing quadruplexes bound to an anthraquinone[79] and to PIPER.[74] Other modelling studies have also been conducted,[68] and images of some of these are shown in figure 1.3.2

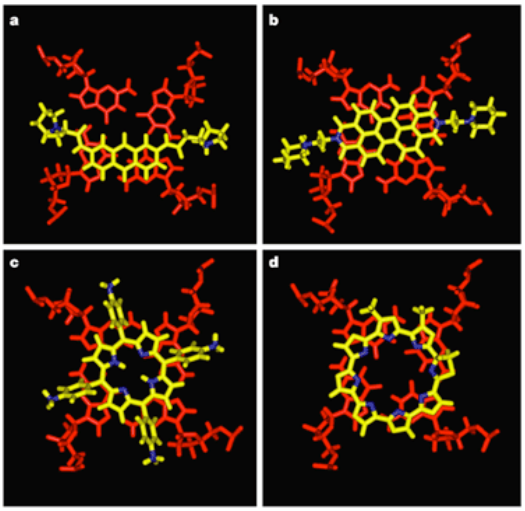


Figure 1.3.2. Modelling results for some small molecule quadruplex binders (yellow) against the top tetrad of a parallel human telomeric quadruplex (red). Ligands are: a) Bi-substituted acridine; b) PIPER; c) TMPyP4; d) Telomestatin. Figure taken from Neidle and Parkinson.[68]

Unfortunately, many quadruplex-binding ligands also bind duplex DNA with relatively high affinities, presenting problems for quadruplex targeting in the presence of duplex DNA. Because there is a large excess of duplex DNA over quadruplex DNA, a very high selectivity indeed would be required to specifically bind quadruplexes. The other problems is that because most ligands target the π-surface of the tetrad stack, they show very little selectivity between quadruplexes formed from different sequences. This could potentially be achieved by targeting the loop sequences or by joining a quadruplex-binding

motif to a ligand specific for an adjacent duplex sequence. Our laboratory is currently investigating both of these approaches.

1.3.2 Proteins that bind quadruplexes

A variety of proteins have also been found to specifically bind quadruplexes. Several of these occur naturally, whereas others have been selected for their binding properties.

The naturally occurring proteins include the helicases implicated in Bloom's[83] (see below)and Werner's[84] syndromes, which have been shown to specifically unwind quadruplexes and to be 'inhibited' by ligands which bind quadruplexes.[85] The *Saccharomyces cerevisiae* protein RAP1[86,87] and the β-subunit of the *Oxytricha nova* telomere binding protein[88] have both been shown to promote quadruplex formation and bind the resulting quadruplexes. The rat hepatocyte protein qTPB42[89] acts to protect quadruplexes from heat denaturation and nuclease digestion. Quadruplexes also interact with various natural enzymes and have been shown to stimulate the activity of a DNA polymerase[90] and inhibit telomerase.[91] The fact that these proteins selectively recognize this non-standard quadruplex motif presumably means that it is capable of forming naturally *in vivo*.

This is perhaps best exemplified by considering Bloom's syndrome, a rare condition characterized by genomic instability and a high level of cancer. It is caused by defects in the *BLM* gene, which encodes a 1417-amino acid helicase. This helicase was shown by Sun *et al.*[83] to have an unwinding activity on quadruplex substrates to yield single strands in an ATP-dependent fashion. It can also unwind duplex DNA, but preferentially acts on quadruplex DNA, as determined by competition experiments and because less of the helicase is required to unwind quadruplex than duplex. Hence it is believed that the *BLM* helicase is required to remove quadruplexes that form prior to recombination and replication. This may explain why it cannot be substituted by other helicases, and hence causes such severe effects. It also implies that quadruplex formation does occur *in vivo*, or else there would be no reason for such a helicase to exist, and its absence would not have significant effects.

An engineered zinc finger protein, Gq1,[92] which specifically binds quadruplexes and inhibits telomerase,[93] has also been developed using phage display (Figure 1.3.3). The template used as a starting point for this was Zif-268,[94] a three zinc finger mouse transcription factor which binds duplex DNA of a specific sequence. Plückthun et al have also developed antibodies raised against quadruplexes and demonstrated that they bind *Stylonychia lemnae* telomeres.[52] This supports the argument that at least part of the telomeres fold into a quadruplex structure.

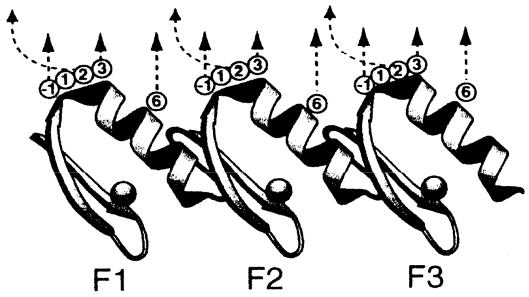


Figure 1.3.3. Structure of Gq1. Gq1 is a three zinc finger protein, and is believed to have the same structure as the parent protein Zif-268, for which the above structure was obtained. Mutated positions are numbered and shown for the three fingers F1, F2 and F3. Zinc atoms are shown at the base of each finger.

1.4 Non-natural quadruplexes

A number of quadruplex sequences have been selected or designed artificially. These include the thrombin binding aptamer (TBA),[95,96] a human immunodeficiency virus (HIV) inhibitor[97] and many simple artificial sequences.[98,99] There has been considerable study of the TBA, which has the sequence d(GGTTGGTGTGGTTGG). Several variant sequences were tested, with modifications in the loops (individually and simultaneously) and the guanine runs.[100] Most of the mutations were destabilising, except for one, which elongated all the guanine runs by one base. Elongating the loops tended to destabilise the structure entropically, and shortening the central loop was destabilising due to steric effects. Removing the terminal guanines was destabilising.

There has also recently been a collection of artificially designed sequences showing non-guanine tetrads,[101-103] interlocked structures,[104] and other unusual features such as intrastrand leaps and chain reversal.[40] Quadruplexes have also been used to form continuous 'G-wire' structures[105] and may function as a nanomotor.[106]

1.5 Natural non-telomeric quadruplexes

The study of natural non-telomeric quadruplexes is the topic of this dissertation. Some work on these structures had been performed previously, and indeed the first report of a biological quadruplex was of a non-telomeric sequence, when quadruplexes were implicated in meiosis.[31] Since then, a few other examples have been described in the literature.

These include the fragile X syndrome repeat $d(CGG)_n$,[84,107,108] the Cystatin B promoter,[109] which has a region with sequence $(CGCG_4CG_4)_4$ and is involved in epilepsy. G-rich strands of the human insulin gene can form quadruplexes,[110] as can the mouse *Ms6-hm* hypervariable satellite repeat,[111] with sequence $(CAGGG)_n$. Of particular relevance in this dissertation is the promoter of *c-myc*,[112,113] which is discussed in detail in §1.6. Hurley has recently proposed that the RET protooncogene can form a quadruplex,[114] and Xodo has proposed that the Ki-ras promoter could form a tetraplex.[115] G-rich RNA can also fold into quadruplex structures, for example the insulin-like growth factor II (IGF II) mRNA.[116]

1.6 C-myc promoter

The *myc* family of oncogenes encodes phosphoproteins that activate other genes downstream that are involved in promoting cell growth. They have been found to be overexpressed in a very wide variety of cancers, including lymphomas and leukaemias, as well as lung, cervical, ovarian, breast and gastric cancers.[117]

Studying *myc* activity is complicated by the fact that it uses four promoters.[118] However, around 80% of the total *c-myc* transcription is controlled by one

particular control region, 115-142 basepairs upstream of the *c-myc* promoter P1.[119] This locus has an unusual homopyrimidine•homopurine (specifically, C•G) arrangement.

It has been demonstrated previously that this region is hypersensitive to nuclease activity, and hence it was named the nuclease-hypersensitive element III_1 (NHE).[120,121] This observation implies that the structure of the DNA in this region is not duplex, but rather more accessible to nuclease activity, which prompted a series of studies to try to block this region. It was found[122] that a purine-rich oligonucleotide d(TGGGGAGGGTGGGGAGGGTGGGGAAGG), complementary to the coding strand of this region, was able to specifically inhibit *c-myc* transcription in HeLa cell lines.[123] It was originally proposed that this oligonucleotide formed a purine•pyrimidine•purine triplex structure with the NHE region, which interfered with the binding of transcription factors.[122] However, such a triplex would require non-physiologically high concentrations of Mg^{2+}, and they have not been observed *in vivo*.[124]

In 1998, Simonsson and coworkers[113] proposed that the purine (and G-rich strand of *c-myc* could in fact form a quadruplex, and demonstrated using DMS footprinting and a polymerase stop assay that it did form a quadruplex *in vitro*. They proposed a model in which the purine-rich strand forms a quadruplex, leaving the pyrimidine-rich strand available for nuclease activity and transcription factor binding, thus resulting in gene activity.

Hurley and coworkers[59] have developed this concept further in a search for therapeutic treatment of *myc*-related cancers, first demonstrating that the small π-rich quadruplex-binding molecule PIPER (§1.3.1) could bind the purine-rich strand of the NHE, and could induce quadruplex formation from the native duplex strand for the NHE, although not for the human telomeric duplex.

In 2002 they demonstrated[125] that a specific G to A transition, which destabilised the formation of a quadruplex from the NHE, resulted in a three-fold increase in transcriptional activity. They also demonstrated that the porphyrin derivative TMPyP4 (§1.3.1) bound this quadruplex and decreased *myc* expression. Taken

together, these results imply that the NHE does indeed form a functionally important quadruplex, but one which acts as a transcriptional repressor, rather than the activator proposed by Simonsson.[113]

Hurley and co-workers then demonstrated[126] that TMPyP4 was capable of inhibiting cancer growth *in vivo*, and that it downregulated both *c-myc* and the human telomerase reverse transcriptase (hTERT). However, it was not clear whether hTERT was independently downregulated, as it is also regulated by *c-myc*.

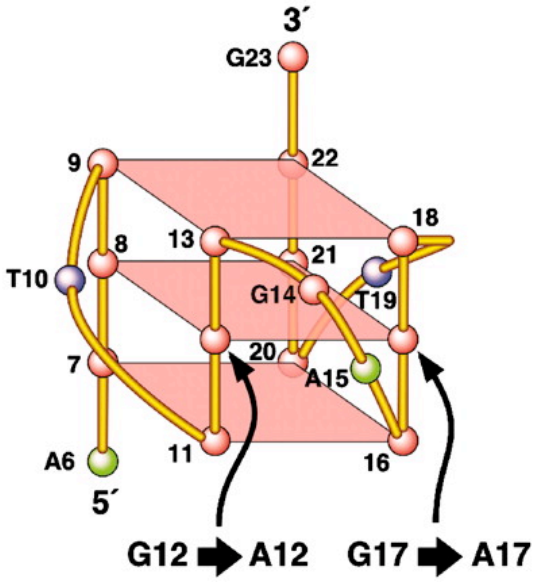


Figure 1.6.1. Proposed structure of the quadruplex formed from the purine-rich strand of the c-myc NHE region. Specific transitions shown to be important for formation are highlighted. The sequence shown is d(AGGGTGGGGAGGGTGGGG), with guanine runs underlined. The runs of four guanines allow for dynamic change, with different sets of three being used in tetrad formation.[127]

Very recently, Hurley and co-workers[127] and Patel and coworkers[128] published adjacent papers examining the structure of the quadruplex formed at the NHE. Hurley performed systematic G to T mutations and molecular modelling, and Patel solved the high resolution solution structure of the complex by NMR spectroscopy. Both concluded that the structure is a parallel quadruplex, (as shown in figure 1.6.1), and that there is a considerable degree of dynamic character, with the loops able to adopt different lengths by using different combinations of three guanines to form the tetrads. This adds considerably to the entropic stability of the folded structure. This structure is in stark contrast to the

antiparallel structures Hurley had previously proposed,[125] although he does suggest that these may still be formed upon drug binding.[127]

1.7 The quadruplex hypothesis

Stimulated by the work describes in § 1.6, I propose the hypothesis that there are a variety of sequences which are important in gene regulation. At its simplest level, this hypothesis proposes that there exist quadruplex-forming sequences in the genome that could act as drug targets for downregulating gene expression, either by acting as a steric block to transcription, or by preventing the binding of components necessary for activation. This feature could be useful therapeutically, if compounds that bind tightly and specifically could be found. (Figure 1.7.1)

It is a generally true observation, if trite, that anything mankind can do, Nature can do better. If we can envisage a method of controlling gene expression *via* a quadruplex mechanism, then it is entirely possible that this is used in at least some instances as a natural control mechanism. Quadruplexes could act as a marker for protein binding, providing an alternative recognition method to purely sequence-based specificity. These could then be used for almost any kind of signalling, from transcription factor binding sites, to splice sites to translation start sites.

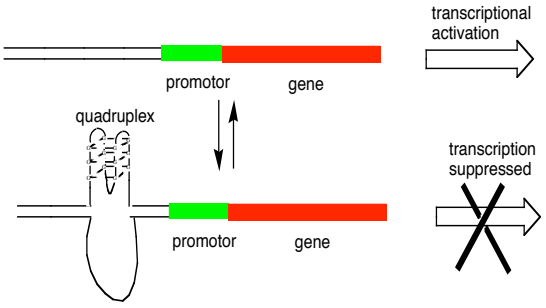


Figure 1.7.1. One version of the quadruplex hypothesis. In the absence of a quadruplex, a gene can be transcriptionally activated. When a quadruplex is formed, this acts as a steric block, either to transcription or to transcription factor binding, resulting in the suppression of transcription. In the presence of a quadruplex-binding drug, this equilibrium favours quadruplex formation, decreasing the level of gene expression.

In this thesis, I investigate this hypothesis. Chapter 2 deals with the question of defining a quadruplex from its sequence alone, which involves a number of

biophysical studies, leading to a ‘folding rule’ for quadruplexes. It then identifies a number of potential candidates for drug targeting in a variety of oncogenes, and demonstrates that they do form quadruplexes *in vitro* that can be targeted using small molecules. An initial *in vivo* test for drug targeting is also described.

Chapter 3 then addresses the idea that quadruplexes could play a natural role. Basic statistics on quadruplexes in the genome are identified, and by considering the structure of the sequences found it is shown there that there may be selective pressures operating to reduce the number of sequences that can form quadruplexes. In order to identify candidates for functional quadruplexes, aspects of genomic information such as cross-species conservation and the location of single nucleotide polymorphisms (SNPs) are considered, leading to the identification of some initial candidates for naturally functional quadruplexes. Chapter 4 reviews this work, and briefly discusses the potential future directions it should take and provides a perspective on the significance of this new field.

CHAPTER 2

G-quadruplex structural studies

2.1 Introduction

There are a number of single-stranded sequences that have been shown to fold into a quadruplex structure under various pH and salt conditions. This includes both physiologically relevant sequences such as the human telomeric repeat sequence $d(GGTTAG)_n$,[27] and artificially-derived sequences such as the thrombin binding aptamer.[96] There is also a selection of techniques that can be used to study the quadruplex structures formed, ranging from simple spectral techniques such as UV and CD spectroscopy[87,129] to high-resolution structural tools such as NMR spectroscopy[26,36] and X-ray diffraction.[22,29]

Despite these previous studies, there has as yet not been a significant effort to address the question as to whether a particular sequence will form a quadruplex under certain conditions, based purely on examining its sequence. In some ways, this could be seen as analogous to the challenge faced in the field of protein folding of predicting protein secondary structure from the primary sequence.[130] In the first half of this chapter, this problem is addressed experimentally to determine the parameters that may be involved, and from them derive a 'folding rule', which defines a set of sequences that are expected to form quadruplex structures under given conditions.

Then, equipped with a rule to identify potential quadruplexes, a selection of quadruplex-forming sequences is identified and characterised in and around the promoter regions of a selection of oncogenes. Using these as examples, their potential as a drug target for the control of gene expression is discussed. Preliminary attempts for *in vivo* analysis of small molecule activity in the fruit fly *Drosophila melanogaster* are also described.

2.2 Development of folding rule

In order to identify and investigate novel quadruplex-forming sequences, it is necessary to develop an understanding of which sequences will form a quadruplex. It is necessary to have an algorithm which will predict secondary structure (quadruplex/non-quadruplex) from the primary sequence.[130] Limited ability to predict the tertiary structure (folding pattern) will also be useful.

There are four aspects that need to be considered, and these are discussed in turn. These are the strand stoichiometry, the number of stacked tetrads in the quadruplex core, the presence of mutations or deletions, and the length and composition of loops. Some of these may be resolved by consideration of the structures and previously known results, others are resolved by experimentation.

It should be emphasised that the rule derived is intended to be sufficient but not necessary; in other words, it is intended to have false negatives (unidentified quadruplexes) but to avoid false positives (incorrectly identified quadruplexes). Hence in developing the rule it is accepted that some possibilities which could lead to quadruplex formation are not considered.

2.2.1 Strand stoichiometry

As discussed previously, quadruplexes can be uni-, bi- or tetramolecular. Except at relatively high concentrations, the unimolecular sequences are favoured.[131] This is as expected from a simple entropic argument. Since under physiological conditions the strand concentration of DNA is relatively low, except in rare exceptions such as *Stylonychia lemnae* macronuclei,[52] interstrand quadruplexes will be strongly disfavoured and for the purposes of the folding rule only intramolecular structures are considered.

2.2.2 Number of tetrads

Structures can form from any number of G-tetrad stacks. In general, the stability increases with increasing numbers of stacks. Single G-tetrads have only been

reported in highly concentrated guanine solutions at mM concentrations,[7,8] and are not physiologically relevant. There are a few examples of double-stack quadruplexes, such as the thrombin-binding aptamer (TBA)[96] and the sequences identified as responsible for the fragile X syndrome.[107,108,132] However, few double-stacks have to date been identified biologically and are in general less stable with regard either to single stranded forms or duplex formation.[100,108] For that reason, only sequences capable of forming three or more G-tetrad stacks have been considered.

2.2.3 Mutations and deletions

Does a quadruplex have to be made up of perfect G-tetrads? Can it tolerate mutations of the guanine bases? Can it tolerate gaps? A few studies have recently been published identifying tetrads not comprised purely of guanine,[133,134] but most of these are artificially designed sequences, where a mixed tetrad is stabilised by flanking G-tetrads – these are not physiologically relevant. In addition, there have recently been structures identified as having 'skipped' structures, where two quadruplexes slide against each other[104] and intrastrand leaps, in which a particular stack of guanines comes from more than one run[40] but these are also artificial structures.

A study has been performed to explore the minimal requirements for quadruplex formation for a simplified variant of the human telomeric quadruplex, and this is described in §2.3. In essence, although some mutations and deletions can be tolerated, they have a significantly destabilising influence, and for that they have been neglected in developing the folding rule.

2.2.4 Loop length and composition

An intramolecular quadruplex must have three loops to link the tetrads together, and they play a large role in determining both the stability and folding pattern of the quadruplex.[23,100,135] If the loops are too short, then it will not be able to form a quadruplex, and if they are too long, then the entropic cost of forming a quadruplex will overcome the enthalpic gain. Hence the optimum stability must lie between these two extremes.[100,135] However, it is not obvious what the limits of loop length must be, or the optimal loop length. Because of this, a project was

conducted systematically investigating this question, and this is described in detail in §2.4.

2.3 Minimal quadruplex studies

2.3.1 Background and rationale

In previous work[93] in our group, studies were performed on a series of oligonucleotides based on the human telomeric repeat, $d(GGTTAG)_n$, with n ranging from 2 to 5. It was anticipated that for n = 2, 3 or 4 no evidence of intramolecular G-quadruplex formation would be found, as there was no set of four d(GGG) runs to form the core of the quadruplex. In contrast, for n = 5 it was expected that an intramolecular quadruplex would be formed.

Name	Sequence	No. GGG	Quad. formation predicted?
n=2	d(GGTTAGGGTTAG)	1	No
n=3	d(GGTTAGGGTTAGGGTTAG)	2	No
n=4	d(GGTTAGGGTTAGGGTTAGGGTTAG)	3	No
n=5	d(GGTTAGGGTTAGGGTTAGGGTTAGGGTTAG)	4	Yes

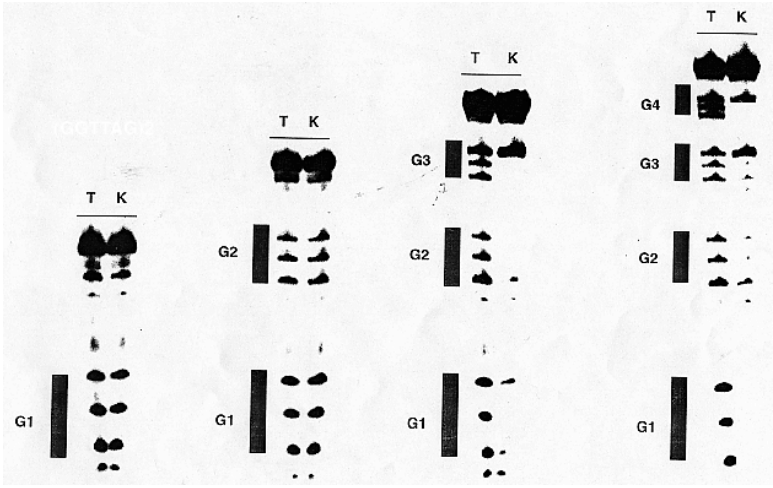
Table 2.3.1. Sequences of $d(GGTTAG)_n$ for n=2 to 5. Each extra repeat introduces an extra run of GGG, and only when n=5 does this reach the value of 4 predicted initially to lead to intrmolecular quadruplex formation.

Two techniques were used to confirm this. The first was a chemical method called dimethylsulphate (DMS) footprinting.[136,137] In this technique, DNA sequences are subjected to the methylating agent DMS. This will methylate the nucleophilic N_7 of guanines, unless they are involved in hydrogen bonding. This methylation renders the guanine available for depurination by water, and the resulting abasic site may then be cleaved with piperidine. If the sequences have been terminally radiolabelled, the fragments may then be separated and identified using polyacrylamide gel electrophoresis (PAGE).

In sequences that form quadruplexes, the N_7 position of the guanines are involved in hydrogen bonding, and so protection from cleavage will be observed.

In contrast, sequences that don't form a quadruplex will show cleavage at every guanine position. In order to distinguish these two alternatives, given sequences are normally prepared under high potassium conditions (to favour quadruplex formation) or low potassium conditions. Thermal denaturation prior to treatment with DMS can also be used to prevent quadruplex formation. In either case, the two separate gel lanes may be compared to identify protection.

The results from the DMS footprinting assay on the sequences d(GGTTAG)$_n$ are shown in figure 2.3.1,[93] with thermally denatured sequences labelled as T, and potassium-rich solutions as K. For n = 2 and 3, no protection is observed, implying that quadruplex formation does not occur even in the presence of potassium. For n=5, as expected, considerable protection is observed in all four sets of guanines (marked with black bar), implying that quadruplex formation does occur in this case. Unexpectedly, for n = 4, protection is also observed, implying that this sequence also forms a quadruplex, despite only having three runs of GGG.



n=2 n=3 n=4 n=5

Figure 2.3.1: PAGE Gel of DMS cuts performed on d(GGTTAG)$_n$ for n=2 to 5. Each sequence was 5'-radiolabelled annealed without (left) or with (right) potassium ions, and treated with dimethyl sulphate (DMS) and piperidine, prior to being subjected to denaturing gel electrophoresis. Bands correspond to cleavage sites. For n = 2 and 3, no protection in potassium is observed. For n=4 and 5, most of the guanines show protection from cleavage, implying that they are involved in quadruplex hydrogen bonding. From S. Patel, PhD thesis, Cambridge Univ. 2000.

Another approach that was used to study these sequences was an enzyme-linked immunosorbent assay (ELISA), using the engineered quadruplex-binding

protein Gq1. In this method (see figure 2.3.2), an indirect measure of binding is used. Biotinylated DNA is affixed to a streptavidin-coated plate and incubated with Gq1, conjugated to a recognition tag. This is then recognised by an antibody, to which is conjugated a reporter enzyme, capable of producing a measurable output. Washing steps are used at each stage to remove non-specific binding interactions.

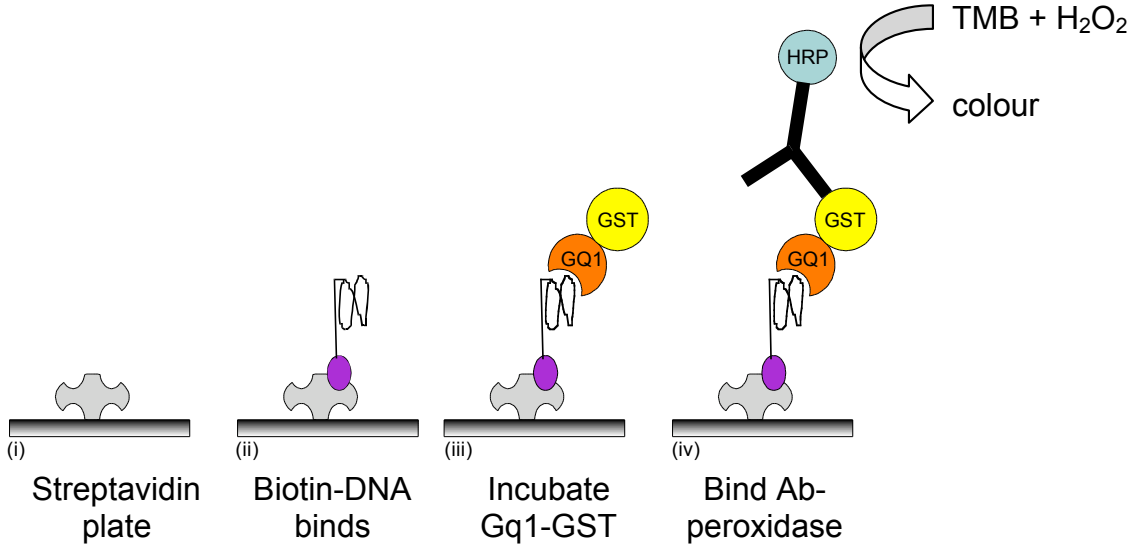


Figure 2.3.2. An ELISA assay. i) A streptavidin plate is prepared. ii) the preannealed quadruplex covalently attached to biotin is incubated with the plate, and any residue washed off. iii) The biotinylated DNA/plate is incubated with the protein Gq1, which is fused to a glutathione sythetase (GST) tag. Any unbound material is washed off. iv) An antibody against the GST tag conjugated to the enzyme horse radish peroxidase (HRP) is added, and any unbound material washed off. Addition of TMB/peroxide gives a coloured output when catalysed by HRP, and this colour may be measured spectrophotometrically to give a measure of the binding between Gq1 and the quadruplex.

When this assay was performed on these samples, variable concentrations of DNA were used to obtain a binding curve. A good binding curve was obtained using $d(GGTTAG)_5$, against which the protein had been selected for binding, no binding was observed for $d(GGTTAG)_2$. For $d(GGTTAG)_3$, binding was only observed at high concentrations, and no good binding curve could be produced. $d(GGTTAG)_4$ behaved very similarly to the longer $d(GGTTAG)_5$, and was bound with a very similar binding constant. This also suggests that it is forming a quadruplex, similar to the longer sequence.

These observations were used as the basis of a study to test the minimal characteristics for quadruplex formation. The first step in this study was to explain the observation that the sequence $d(GGTTAG)_4$ was apparently capable of forming a quadruplex. Systematic mutations and deletions were then used to investigate the thermodynamic contributions of each base in the tetrad stack.

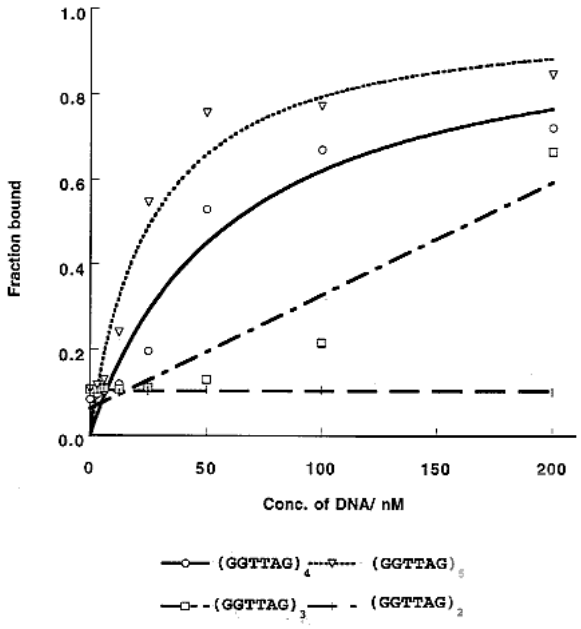


Figure 2.3.3: ELISA binding assays to measure the binding of the engineered zinc finger quadruplex-binding protein Gq1 on a series of human telomeric sequences. Good binding curves are observed for n = 4 and 5, but not for n = 2 and 3. From Patel, PhD thesis, 2000

As well as being useful for pure structural studies, there may be biological and pharmaceutical relevance to having minimally stable G-quadruplex structures as screens for quadruplex-binding agents. This is because a less stable initial structure would make the stability change available from small molecule binders more significant and easier to examine rapidly.

2.3.2 Structural hypothesis

Although all published quadruplex structures solved when this work was performed had the 5' and 3' ends of the sequence not being part of the G-tetrad core of the quadruplex, it was hypothesised that the structure for $d(GGTTAG)_4$ (**QL1** – see table 2.3.2) could be an intramolecular G-quadruplex with the 5' and 3' ends internal to the structure (Figure 2.3.4). Since this hypothesis was proposed, a similar structure has been proposed for an intramolecular structure

based on the *i*-motif and two bimolecular G-quadruplex structures.[138] The structure shown is an antiparallel form, based on the NMR structure solved by Wang and Patel.[27] An alternative structure would involve a structure based on the parallel crystal structure solved by Neidle and co-workers.[29]. The two structures can be resolved by circular dichroism spectroscopy, as described in section 2.3.4.

i) Three levels of G quadruplex

ii) Two levels of quadruplex, stacked triplex

Figure 2.3.4. i) proposed structure for **QL1**, with terminal guanines involved in tetrad formation ii) Alternative equilibrium form of **QL1**, with the 3' guanine unbound to relieve entropic constraints on the loop.

If this hypothesis were correct, then deletion of the terminal guanine bases would be expected to disrupt the structure, in contrast to free ends, where deletion generally has little thermodynamic effect. In addition, it might be expected that deletion of the 3' G would have relatively little effect compared to the deletion of the 5' G's, as the entropic cost of immobilising that G and constraining the 3-base loop leading up to it would be large compared to the enthalpic gain from only one base binding to the quadruplex.

To confirm the hypothesis, parent sequence **QL1** and mutants **QL2** and **QL3** were studied using UV melting.[129] The results and sequences are shown below. Little evidence of hysteresis between the melting and annealing runs was observed, and the results from heating and cooling profiles were very similar, supporting the suggestion that these are equilibrium experiments.

Oligo	Sequence	Notes	T_m (/°C)
QL1	GGTTAGGGTTAGGGTTAGGGTTAG	Parent	62.3
QL2	--TTAGGGTTAGGGTTAGGGTTA-	No terminal G's	N/A
QL3	GGTTAGGGTTAGGGTTAGGGTTA-	No 3'G	59.1

Table 2.3.2: Parent sequence **QL1** and mutant forms **QL2** and **QL3** were subjected to UV melting between 10 and 90°C at 5 μM concentration in GQ buffer. UV measurements were performed at 295 nm and corrected for buffer absorption. Van't Hoff analysis was performed to obtain T_m values. **QL2** showed no melting transition. Values are the average of three repeats. Errors are estimated as ±1 °C

These results show that **QL1** and **QL3** formed quadruplex-like structures, with T_ms of 62 and 59 °C respectively. **QL2** did not show a melting transition at 295 nm, consistent with it not forming a quadruplex-like structure. This is supportive of the initial hypothesis, that the terminal G's were directly involved with quadruplex formation, and also that the 3' G is less involved, as its deletion does not affect the quadruplex stability substantially.

2.3.3 Mutational experiments

Having established that imperfections can be tolerated within the quadruplex template, it was decided to proceed with a more complete study of the minimal requirements for quadruplex formation by systematic replacement of the G's in one strand of the quadruplex (positions X, Y and Z in figure 2.3.5), with either gaps (base deletions) or adenine. Adenine was selected due to the similarity of its π-surface with guanine. These oligonucleotides were studied by UV melting and van't Hoff analysis as discussed previously.

Disruption, either by deletion or substitution, of the X position (**QL3**, **QL7**) is not especially destabilising to the structure, with only a small ΔT_m observed. In contrast, disruption of either the Y or Z positions (**QL5**, **QL6**, **QL8**) seems to be heavily destabilising. This can be explained in terms of the equilibrium mentioned previously – the X position is only relatively loosely bound for entropic reasons, whereas the Y and Z positions are more firmly bound. Hence disruption in these latter positions is more destabilising. This can be seen as further evidence for the initial structural hypothesis.

The other observed feature is that in every case, replacement of a gap with an adenine is destabilising, in contrast to what may be expected from a π-stacking model. However, adenine does break up the H-bonding network, losing three H-bonds and a carbonyl-cation interaction as well as increasing the steric bulk, and so this could have a large effect. Inosine would have a less disruptive effect, losing only one H-bond, but this is of limited biological relevance.

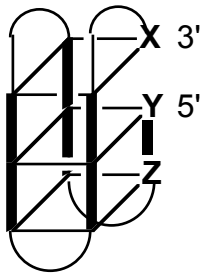


Figure 2.3.5: Generic template for minimal structure examination. X, Y and Z refer to the positions to be altered; the full sequence would be d(**YZ**TTAGGGTTAGGGTTAGGGTTA**X**)

Oligo	Sequence	X YZ	T_m (/°C)	ΔT_m (/°C)
QL1	GGTTAGGGTTAGGGTTAGGGTTAG	G GG	62.3	0
QL2	TTAGGGTTAGGGTTAGGGTTA	– ––	N/A	-
QL3	GGTTAGGGTTAGGGTTAGGGTTA	– GG	59.1	-3.2
QL5	GTTAGGGTTAGGGTTAGGGTTAG	G –G	26.0	-36.3
QL6	AGTTAGGGTTAGGGTTAGGGTTAG	G AG	18.5	-43.8
QL7	GGTTAGGGTTAGGGTTAGGGTTAA	A GG	56.3	-6.0
QL8	GATTAGGGTTAGGGTTAGGGTTAG	G GA	18.0	-44.3

Table 2.3.3. Parent sequence **QL1**, deletion mutants **QL2-5** and adenine scan mutants **QL6-8** were subjected to UV melting between 10 and 90°C at 5 μM concentration in GQ buffer. UV measurements were performed at 295 nm and corrected for buffer absorption. Van't Hoff analysis was performed to obtain T_m values. **QL2** showed no melting transition. Values are the average of three repeats. ΔT_m values are shown relative to the parent sequence.

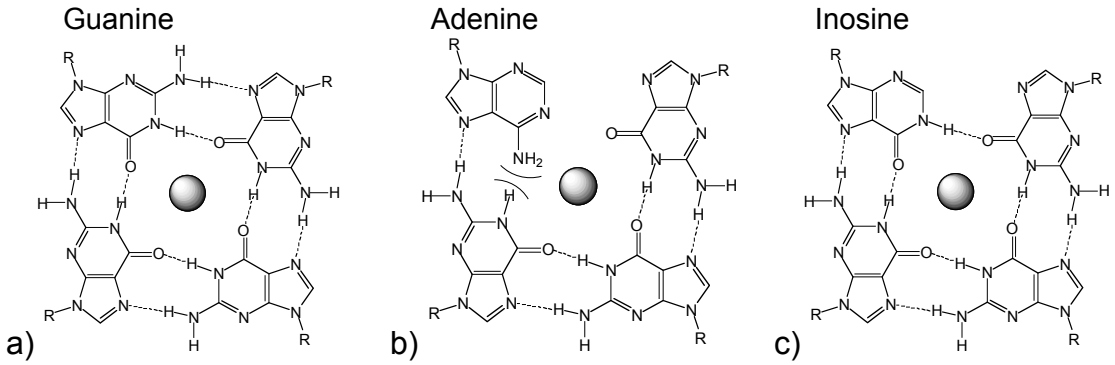
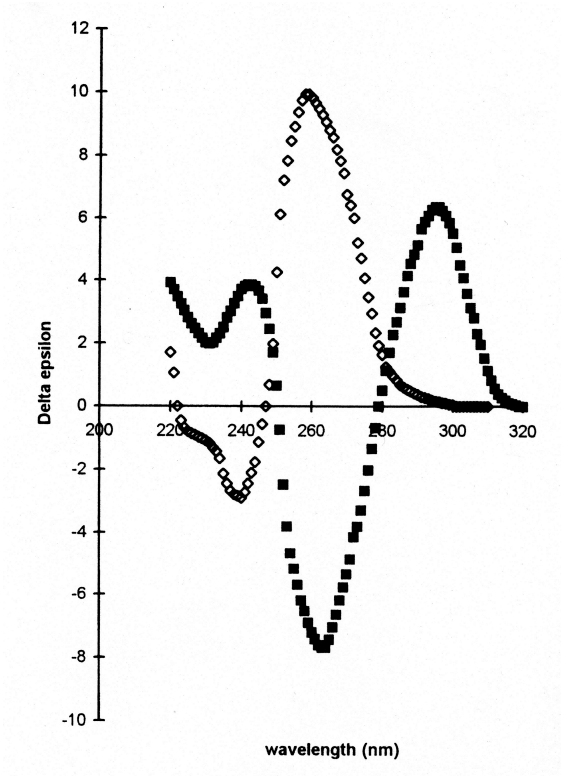


Figure 2.3.6: Structures of tetrads with various mutations, all shown in the top left corner. a) Guanine makes a symmetric hydrogen bonded network b) Replacement of guanine with adenine in a G4 tetrad disrupts the H-bonding in the tetrad considerably, breaking three hydrogen bonds and creating a steric clash. c) Replacement with inosine is much less disruptive, removing only one hydrogen bond.

2.3.4 CD spectroscopy

Circular dichroism (CD) spectroscopy has been used extensively in characterising quadruplex structures[64,87] and other secondary forms of DNA,[139] as well as for protein structural determination.[130,140] It relies on the property of chiral molecules to absorb the two 'stereoisomers' of circularly polarised light differently. In quadruplexes, the key absorption arises from the G-G stacking interaction,[141,142] the strength of which depends on the conformation of the guanines. In parallel quadruplexes the glycosidic bonds are all in the favoured *anti* conformation, but in order to form an antiparallel quadruplex, some have to adopt the *syn* conformation in order for the tetrads to be planar.[143]

As a result of this difference, the absorption profile is different between the two forms. Parallel DNA exhibits a peak around 260 nm and a trough around 240 nm, whereas antiparallel DNA has a peak around 295 nm and a trough around 260 nm.[64,87] The details of these spectra vary with the exact sequences, but these gross features are highly reproducible. An example spectra from Balagurumoorthy *et al*.[64] is shown in figure 2.3.7



2.3.7. Prototype CD spectra. Diamonds show the trace for a parallel quadruplex formed from d(G_{12}), and shows a characteristic peak at 260 nm and a smaller trough at 240 nm. Filled in squares show the trace from an antiparallel quadruplex formed from d($G_4T_4G_4$), with a peak at 295 nm and a trough at 260 nm. Taken from Balagurumoorthy *et al*.[64]

In order to confirm that the secondary structures observed by UV spectroscopy were indeed quadruplexes, and also to determine whether the structures formed were parallel or antiparallel, a series of CD studies were performed on the **QL** sequences. These were studied at 4 μM concentration in GQ buffer at 4°C, having been annealed and slow-cooled over a period of 11h. The results are shown in figure 2.3.8. It may be clearly seen that the sequences studied fall into two categories. Samples **QL1, QL3** and **QL7** show a large peak at 295 nm, whereas samples **QL2, QL5, QL6** and **QL8** exhibit a much smaller peak at 295 nm. All of these are consistent with the formation of an antiparallel quadruplex structure, as shown in the diagrams in the previous sections.

Comparing these results to those obtained by UV spectroscopy (table 2.3.2), it can be seen that the results agree qualitatively, with sequences **QL1, QL3** and **QL7** having relatively high melting temperatures (over 55 °C), whereas samples **QL2, QL5, QL6** and **QL8** have relatively low temperatures (below 30 °C). These low T_m samples may not have full quadruplex formation at temperatures relatively

close to the T_m. Similarly, although **QL2** did not exhibit a clear melting transition in the UV melting experiments, it may have begun the melting transition by 4 °C, accounting for the existence of a CD trace consistent with a quadruplex.

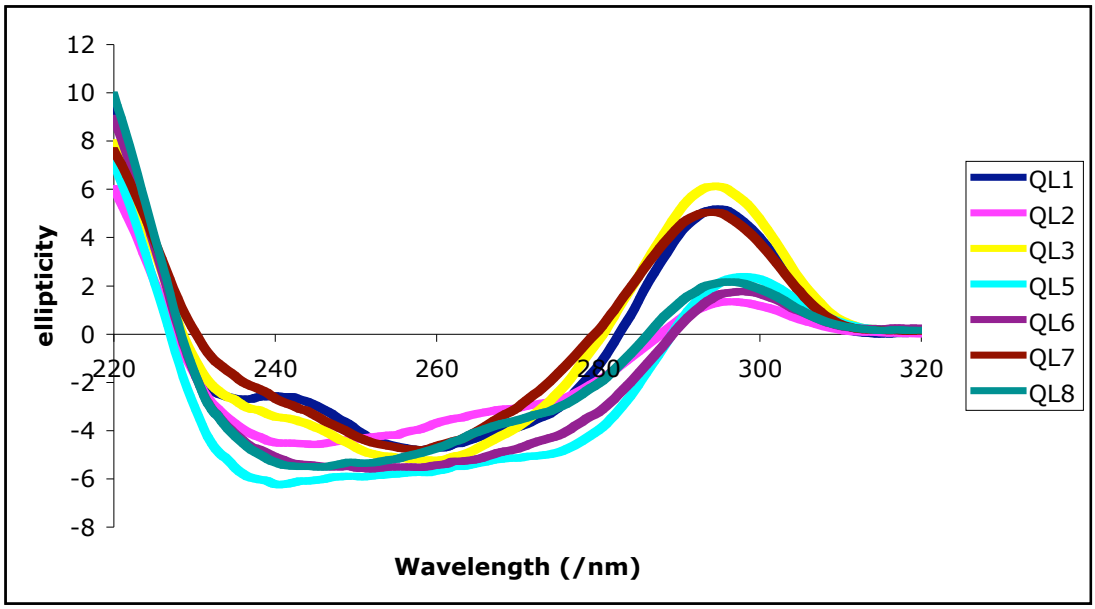


Figure 2.3.8. CD spectra for the **QL** series of oligonucleotides. Data was collected at 4 μM concentration at 4 °C. Samples **QL1, QL3** and **QL7** show a large peak at 295 nm, whereas samples **QL2, QL5, QL6** and **QL8** exhibit a much smaller peak at 295 nm.

2.3.5 Conclusions

The novel structure proposed for **QL1**, with the 3' and 5' ends internal to the tetrad stack is supported by these mutational experiments, as mutation of either terminus has a significant impact on the stability of the formed product. The suggestion that this structure could exist in dynamic equilibrium is also supported, with mutations of the single 3' guanine being significantly less significant in terms of stability than the 5' guanines. The structure appears to adopt an antiparallel arrangement, as found by Wang and Patel for the parent sequence in solution.[27]

Mutating the positions of the tetrad stack to adenine has a significant effect on the stability of the quadruplex, with the exception of the lone 3' guanine, which is an unusual feature in most quadruplexes. Deletions are also highly destabilising. In each case, destabilising effects result in a melting temperature decrease of

over 40 °C, bringing the melting temperature to below physiologically relevant temperatures and reducing the intensity of the CD signal at 4 °C.

Hence it is believed to be valid for the purposes of developing the folding rule to neglect sequences without all the components for a full stack of G-tetrads. While other sequences can form under specialised circumstances, they are less stable and less likely to be relevant physiologically.

2.4 Loop length studies

2.4.1 Introduction

In order to understand how quadruplex folding is affected by loop length, a series of model quadruplexes were studied using both biophysical methods and molecular modelling. The model quadruplex structures were based on the human telomeric repeat, and the two structures that have been solved for it, one parallel (crystal structure in K^+ solution),[29] and one antiparallel (NMR structure in Na^+ solution).[27] This work was carried out in collaboration with Pascale Hazel, who performed all the molecular modelling work described in this section. This work has now been published.[144]

We selected sequences with purely thymine loops for simplicity. Thymine residues within the loops have the capacity to form stacking interactions, and to hydrogen bond with other residues. They are also small enough, compared to adenine, to have several residues in very close proximity in the loops.

Molecular dynamics simulations are now commonly used to obtain structural information about biomolecules. Molecular dynamics simulations of quadruplexes give stable trajectories over nanosecond timescales.[145-149] Free energy calculations using the Molecular Mechanics/Poisson Boltzmann Surface Area (MM-PBSA) post-processing method, in combination with the Amber force field, have been used to look at A and B-DNA conformations[150], RNA loop structures[151] as well as protein folding.[152] Pathways along which four guanine strands might assemble into a quadruplex were suggested based on molecular

dynamics and free energy calculations using the MM-PBSA method.[146] Stefl et al found that the inclusion of quadruplex channel ions was essential in order to obtain meaningful results in the free energy calculations. The free energy calculation method used only yields estimates of the energy of each conformation. Such flexible systems as quadruplexes with loops of up to 6 residues would require extremely long simulations to sample all of their configurational space, and hence any energy calculation must be approximate. However, it was hoped that the free energies would be able to give insight into favoured conformations of quadruplexes.

Oligo name	Oligonucleotide sequence
G_3T	d(TGGGTGGGTGGGTGGGT)
G_3T_2	d(TGGGTTGGGTTGGGTTGGGT)
G_3T_3	d(TGGGTTTGGGTTTGGGTTTGGGT)
G_3T_4	d(TGGGTTTTGGGTTTTGGGTTTTGGGT)
G_3T_5	d(TGGGTTTTTGGGTTTTTGGGTTTTTGGGT)
G_3T_6	d(TGGGTTTTTTGGGTTTTTTGGGTTTTTTGGGT)
G_3T_7	d(TGGGTTTTTTTGGGTTTTTTTGGGTTTTTTTGGGT)
TTA-G_3TTA	d(AGGGTTAGGGTTAGGGTTAGGGT)
TTA-G_3T_1	d(AGGGTTAGGGTGGGTTAGGGT)
TTA-G_3T_2	d(AGGGTTAGGGTTGGGTTAGGGT)
TTA-G_3T_3	d(AGGGTTAGGGTTTGGGTTAGGGT)
TTA-G_3T_4	d(AGGGTTAGGGTTTTGGGTTAGGGT)
TTA-G_3T_5	d(AGGGTTAGGGTTTTTGGGTTAGGGT)
TTA-G_3T_6	d(AGGGTTAGGGTTTTTTGGGTTAGGGT)
TTA-G_3T_7	d(AGGGTTAGGGTTTTTTTGGGTTAGGGT)

Table 2.4.1. Sequences of oligonucleotides used in this study. The first series, G_3T_x, have all loops consisting of x thymines. The second series, TTA-G_3T_x have the first and third loops with the sequence TTA, and the central loop consisting of x thymines. TTA-G_3TTA is the parent sequence, the human telomeric repeat.

There are as yet no guidelines about the specific folded structure adopted by quadruplexes of a given sequence. Structural information is necessary in the rational design of quadruplex-binding drugs. Indeed, different folds generate distinct binding regions for drugs, with very different recognition characteristics. The π-surfaces on the top and bottom of the quadruplex stack may be available for binding,[29] or may be covered by loops.[27] The grooves on the side of the tetrad stacks may be available for binding,[27] or blocked by the loops.[29] Their width is also determined partly by the folding pattern.[27] The loops themselves present a possible target, and their folding structure will determine how they could be bound.

2.4.2 Results

2.4.2.1 Stability of quadruplexes with oligo-dT loops.

A series of oligonucleotides consisting solely of d(GGG) repeats separated by loops consisting of $d(T)_n$ were synthesised and examined (See Table 2.4.1, top half). These G_3T_n sequences were shown to form quadruplexes using UV melting and CD spectroscopy.

UV melting studies on oligo-dT loop sequences. It has been previously shown[129] that quadruplexes melt with a significant hyperchromic shift at 295 nm. The resulting melting curves can be analysed either by finding a maximum in the first derivative, yielding the melting temperature, T_m, or more precisely using a van't Hoff analysis, fitting the melting curve to a two-state model. This latter can be used to yield full thermodynamic parameters as well as the T_m. The $d(T)_n$ loops studied here showed clear melting transitions at 295 nm, with the exception of G_3T and G_3T_7, where any transition was either at too low a temperature to be observed or of too small a magnitude. It was indicated that the problem was not an excessively high T_m by rerunning the samples in low-salt solutions (20 mM KCl), which would reduce the T_m.[129] Under these conditions, G_3T showed a melting transition above 80 °C, whereas G_3T_7 showed no transition. The other sequences had melting temperatures that decreased with increasing loop length, indicating decreased stability with respect to the unfolded structures. This may be explained by considering the increased entropic cost of constraining the longer

loops. Melting temperature data is shown in Table 2.4.2, and sample melting curves are shown in Figure 2.4.1.

Oligo name	T_m (/°C)
G_3T	–
G_3T_2	71
G_3T_3	58
G_3T_4	48
G_3T_5	34
G_3T_6	20
G_3T_7	–

Table 2.4.2. UV melting temperatures for the oligonucleotide sequences with all three loop lengths varying. T_m values are derived by Van't Hoff analysis and have an associated error of ±2 °C. Sequences G_3T and G_3T_7 did not show a melting transition.

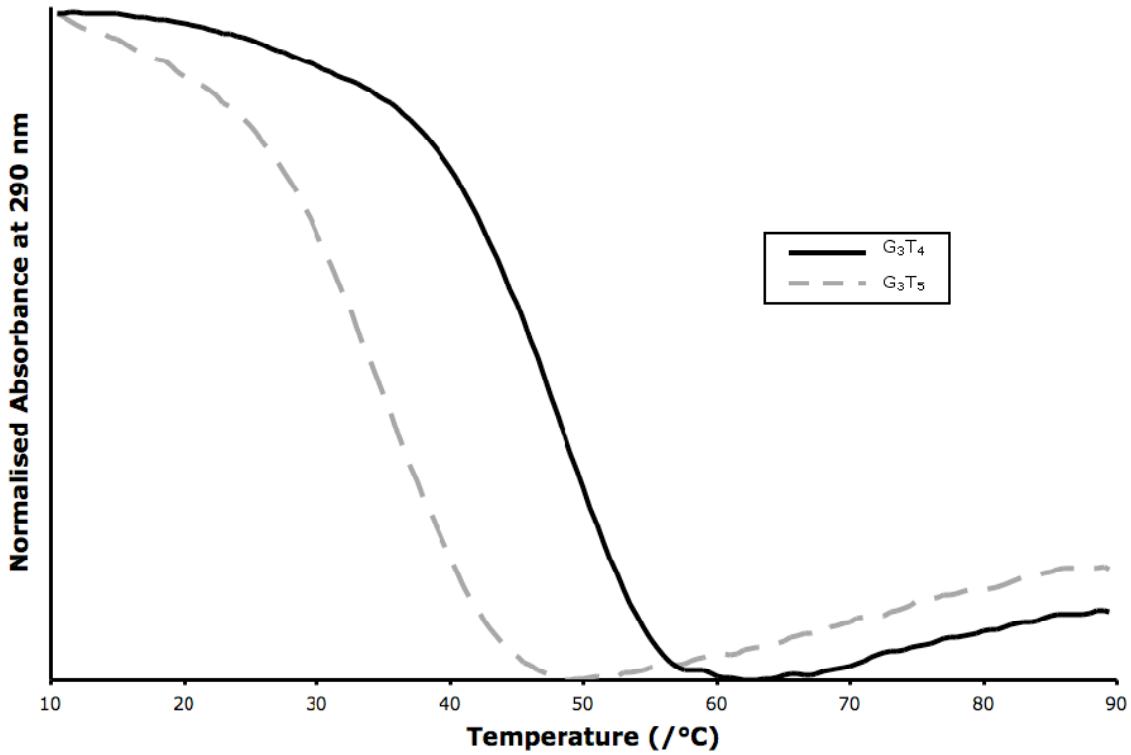


Figure 2.4.1: Sample UV melting curves for sequences G_3T_4 (solid line) and G_3T_5 (hashed line). These results are the average of a series of three heating and cooling cycles between 10 and 90°C.

CD Spectroscopy on oligo-dT loop sequences. Quadruplexes exhibit characteristic circular dichroism signals, depending on which of two broad classes of fold they belong to.[64,87,139] Antiparallel folds exhibit a positive CD signal at around 295 nm, with a negative signal at around 260 nm. On the other hand, parallel folds have a positive signal at approximately 260 nm, and a negative signal near 240 nm. Unstructured sequences exhibit neither of these characteristics. The data shown in Figure 2.4.2 implies that the sequence G_3T is purely parallel, G_3T_3-G_3T_7 are antiparallel, and G_3T_2 exhibits substantial polymorphism, and appears to consist of a superposition of the two alternative states. The decreasing peak at 295 nm with increasing loop length can be attributed to the fact that the long loops behave largely as single strands, and oligo-dT exhibits a small negative CD signal at 295 nm, superimposed on the CD signal resulting from the tetrad core.

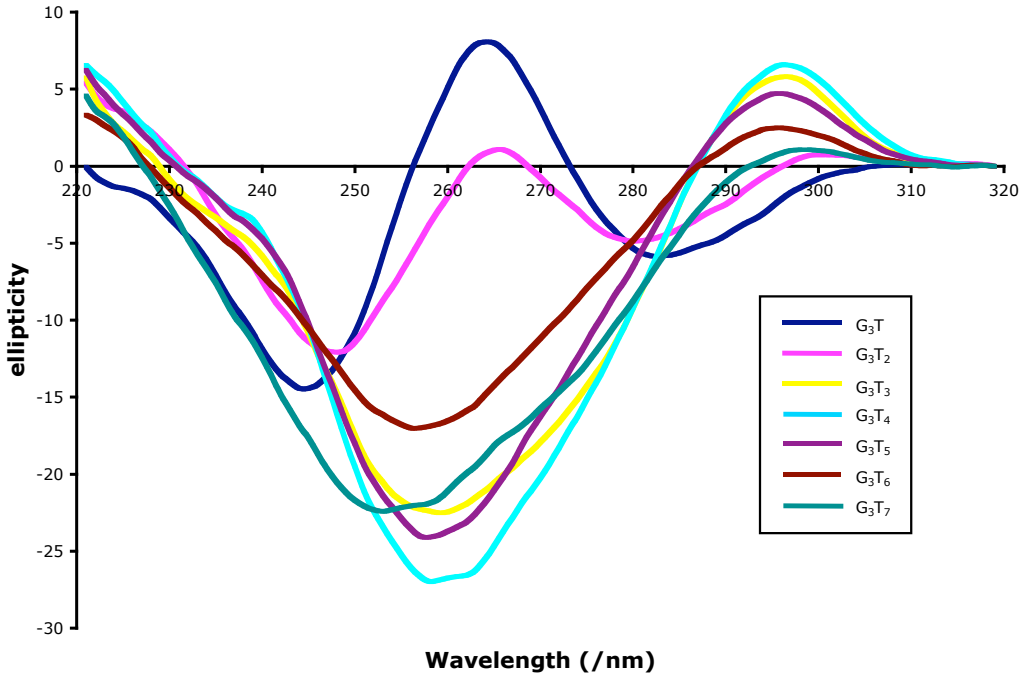


Figure 2.4.2: CD spectra for the oligonucleotide sequences with all three loop lengths varying. Data was collected at 4 μM concentration at 4 °C. Samples G_3T and G_3T_2 exhibit a peak at 265 nm characteristic of parallel folds (G_3T_2 possibly being polymorphic); the other sequences have a peak at 295 nm, characteristic of antiparallel folding.

2.4.2.2 Stability of quadruplexes with two TTA loops and a central oligo-dT loop.

In order to study sequences more similar to the well-characterised human telomeric sequence of $(TTAGGG)_n$, facilitating accurate molecular modelling, a second series of DNA oligomers was studied using the same techniques. The parent sequence TTA-G_3TTA, which consists only of unmodified telomeric repeats, was also studied. These are all shown in Table 2.4.1 (bottom half).

UV melting studies on single-loop variants. This series of sequences showed little difference in the melting temperature, with all showing a hyperchromic transition at 295 nm at around 55 °C. This suggests that these sequences all form quadruplexes under these conditions.

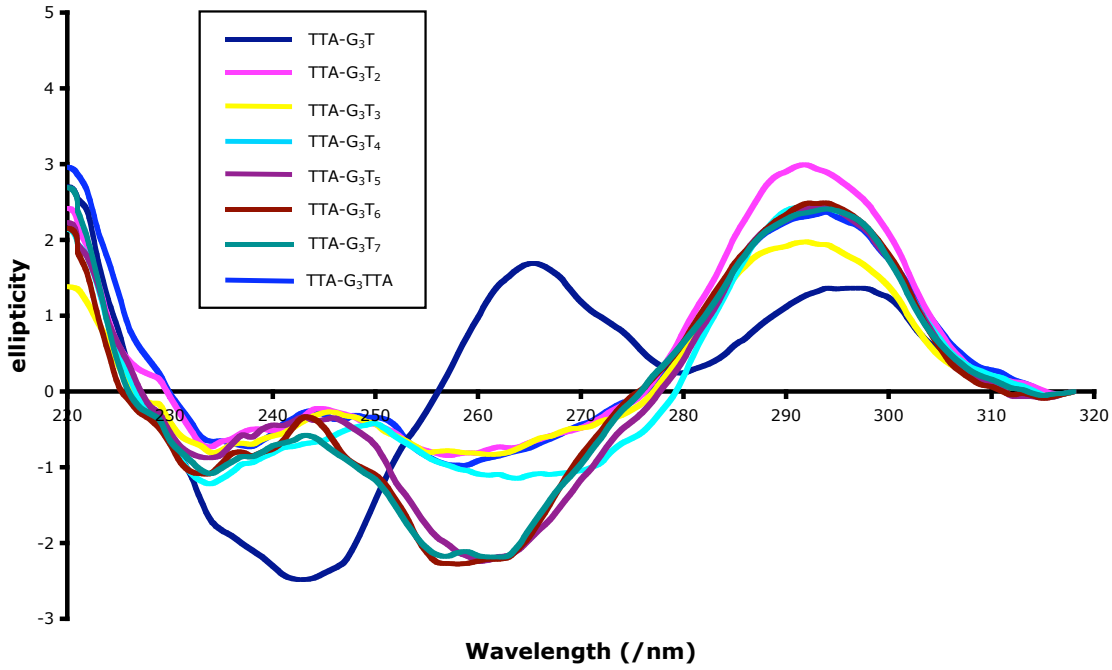


Figure 2.4.3: CD data for sequences with only the central loop varying. Data was collected at 1 μM concentration at 4°C. Sample TTA-G_3T exhibits a peak at 265 nm characteristic of parallel folds (though possibly being polymorphic); the other sequences have a peak at 295 nm, characteristic of antiparallel folding.

CD spectroscopy on single-loop variants. These samples all showed CD spectra consistent with antiparallel folds, with the exception of TTA-G_3T, which had a polymorphic spectrum very similar to that for G_3T_2, suggesting that an

extremely short loop can be accommodated in an antiparallel fold as long as there is only one such loop.

2.4.2.3 Molecular Dynamics simulations.

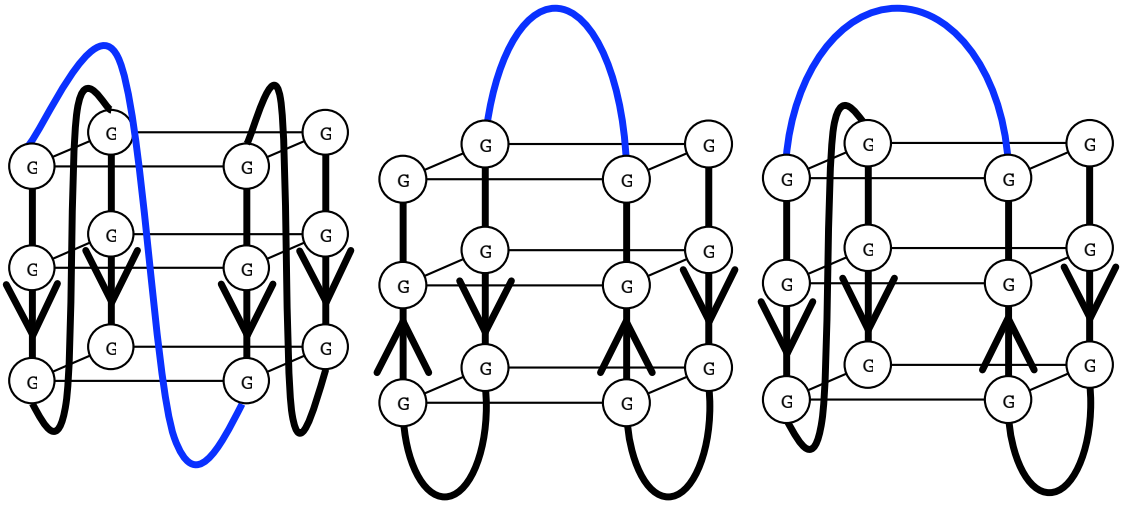
Three template quadruplex structures were obtained from the Protein Data Bank for use in the simulations, to represent the parallel folding conformation and two different types of antiparallel fold. The parallel conformation was based on the human telomeric repeat d[$AG_3(T_2AG_3)_3$] crystal structure (PDB code 1KF1). This structure contains three T_2A strand-reversal loops. For most purposes, the NMR structure of the above human sequence was used as an antiparallel template (PDB entry 143D). This contains two lateral loops and a central diagonal loop. Both these structures were used directly after modification of the central loop to the desired length, ranging from T to T_6.

Sequence	parallel	antiparallel (143D)	antiparallel (186D)
TTA-G_3T	4 ns		4 ns
TTA-G_3T_2	4 ns	2 ns (unstable)	4 ns
TTA-G_3T_3	4 ns	4 ns	
TTA-G_3T_4	4 ns	4 ns	
TTA-G_3T_5	4 ns	4 ns	
TTA-G_3T_6	4 ns	4 ns	
G_3T	4 ns		4 ns (unstable)
G_3T_2	4 ns		4 ns

Table 2.4.3 Simulation periods for molecular dynamics. 143D refers to the antiparallel quadruplex built from the 143D PDB structure, with 1 lateral, 1 diagonal and a final lateral loop. 186D refers to the antiparallel quadruplex built from the 186D PDB structure, with two lateral and one strand-reversal loop. Sequences were the same as those in Table 2.4.1, except that the 3' terminal T was removed, as it is not present in the experimentally derived structures.

For short loop lengths of one or two nucleotides, an alternative antiparallel template with a lateral central loop was derived from the NMR structure of d$(T_2G_4)_4$ (PDB code 186D). This sequence differs from the human sequence by one G to A mutation in each repeat, and contains a mixture of parallel and antiparallel guanine strands, with two lateral loops and a third strand-reversal

loop. The G-quartets were left unchanged, the first and third loops were modified to T_2A, and the central loop modified as described previously. The sequences simulated are shown in Table 2.4.3, and the template structures in Figure 2.4.4.



1KF1 Parallel template

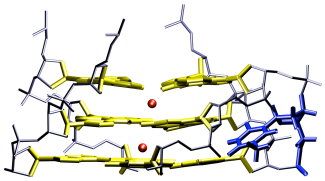
143D antiparallel template

186D antiparallel template

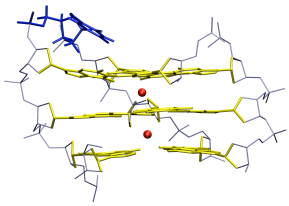
Figure 2.4.4: template structures used in the molecular dynamics simulations. 5' to 3' directions are shown by arrows. The central loop is coloured blue in each case.

Relative stabilities of the simulated quadruplexes were compared using analysis of various criteria, such as root mean square (rms) deviations, planarity and hydrogen bonding integrity of the G-quartets. Loop structures were assessed using stacking and hydrogen bonding abilities as well as rms deviations. The presence of ions within the central core of the quadruplexes was also related to stability. The average structure of each quadruplex over the last 2 ns of simulation is shown in figure 2.4.5.

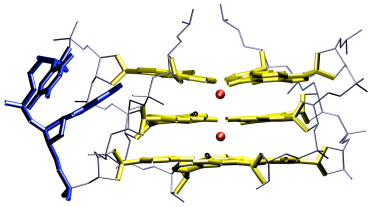
Molecular dynamics simulations of single-loop variants in parallel conformations. All the parallel quadruplex simulations gave stable trajectories for the G-quartets, with low rms deviations between 1.4 and 1.5 Å on average. Different loop lengths did not influence G-quartet stability. The strand-reversal loops showed a much greater range of rms deviations than the G-quartets. The T_2A loop rms deviations were variable between simulations, ranging from 2.4 to 6.1 Å, despite being in identical starting conformations. These T_2A loops are stabilised by base-stacking interactions only, with no hydrogen bond interactions, and have high flexibility in simulated solution.



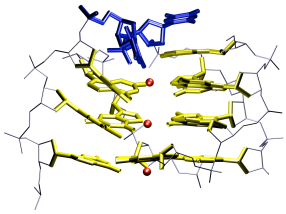
a) TTA-G_3T parallel



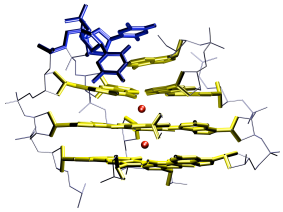
b) antiparallel (186D)



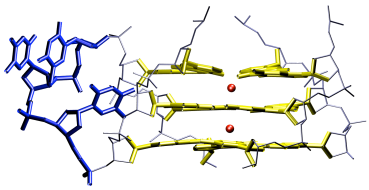
c) TTA-G_3T_2 parallel



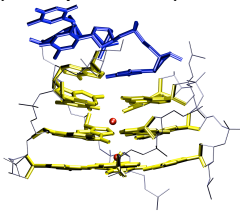
d) antiparallel (143D)



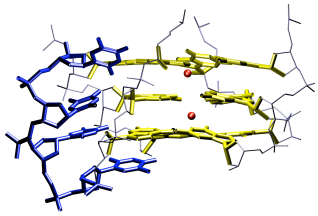
e) antiparallel (186D)



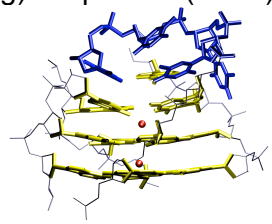
f) TTA-G_3T_3 parallel



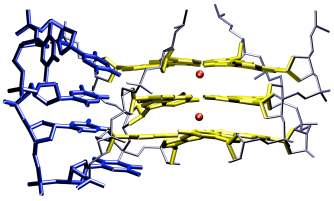
g) antiparallel (143D)



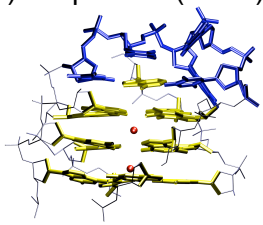
h) TTA-G_3T_4 parallel



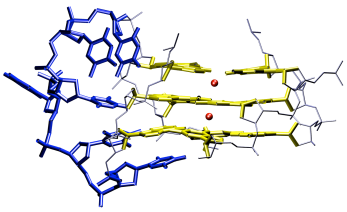
i) antiparallel (143D)



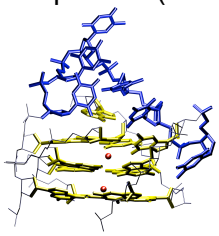
j) TTA-G_3T_5 parallel



k) antiparallel (143D)



l) TTA-G_3T_6 parallel



m) antiparallel (143D)

Figure 2.4.5: Structures of each single-loop variant sequence simulated. The structures were averaged over the last 2ns of dynamics. For clarity only the G-quartets (yellow) and modified loop (blue) are shown. The K^+ ions within the central channel are shown in red.

The parallel quadruplex x-ray structure loop conformation, in which the A residue stacks with the first T, and the second T faces the solvent, was often not retained during the simulations. The loops tended to adopt a less compact conformation, in which the second T residue reached further into the solvent. This T occasionally stacked with the A residue, although only for a few hundred picoseconds at most. At all other times, the A and first T loop residues stacked with each other. Only rarely during any of the simulations did no stacking take place between any two of the loop residues. The T_2A loop flexibility was not linked to the modified loop deviations. The same T_2A loop behaviour was observed during simulations of the human sequence with three unmodified loops.

The modified strand-reversal loops also displayed a range of rms deviations, depending on length, but also on the interactions of the loop residues with the G-quartets. Increasing the loop length from T to T_3 increased the rms deviation of each loop, related to the increasing degrees of freedom. The T loop residue pointed towards the G-quartets, forming stable hydrogen bonds with two guanine residue N^2 atoms. The T_2 loop residues were oriented towards the solvent, but maintained stacking interactions with each other during the whole 4 ns simulation.

The T_3 loop behaved similarly to the T_2A loops during the first 3 ns of simulation, forming alternate stacking arrangements between two of the three residues. However, all stacking between the residues was lost during the final 1 ns of simulation, during which no interactions were present either between the loop residues or between the loop residues and G-quartets. The fact that no stable structure for this loop was achieved during the simulation is reflected in the higher rms deviation of 2.2 Å with respect to the average structure over the last 2 ns of simulation. The T_3 loop rms deviation did not stabilise over the course of the simulation. This instability is not necessarily linked to the nature of the T_3 loop, as periods during which there were no interactions between the T_2A loop residues occurred frequently during the simulations.

The greater length of the T_4 to T_6 loops enabled the residues to point inwards towards the G-quartet, thus forming stable hydrogen-bonding interactions,

resulting in a lower rms deviation for the T_4 loop. The thymine residues were also able to stack with each other, thus further stabilising the loop. The T_5 and T_6 loops had higher rms deviations as some residues were more exposed to the solvent and unable to form stabilising interactions with the G-quartet.

Molecular dynamics simulations of single-loop variants in antiparallel conformations. The single-loop variants also formed stable antiparallel quadruplexes. The T and T_2 loops were too short to span the diagonal of the G-quartets in the 143D antiparallel structure, as shown by the distortion of the TTA-G_3T_2 143D antiparallel quadruplex after 2 ns of simulation (Table 2.4.3). One of the G strands was displaced from the G-tetrad plane, and lost hydrogen bonding interactions with the adjacent G residues. This affected all three G-tetrads. As a consequence, the K^+ ions moved from between to within the G-quartet planes.

Both TTA-G_3T_1 and TTA-G_3T_2 were therefore simulated using the 186D template, yielding stable and undistorted structures over 4 ns. The T_1 loop residue was able to stack on top of one G, and form a stabilising hydrogen bond with the terminal A residue. One of the T_2 loop residues stacked on top of the terminal A residue, but in this case no hydrogen bonds could form. Both loops were stable, as shown by low rms deviations.

The longer loop lengths formed stable quadruplexes in the 143D antiparallel conformation. The distorted 143D T_2 quadruplex retained three K^+ ions within its channel during the 2 ns of simulation. However, all the other antiparallel quadruplexes retained only two ions within the core channel, the K^+ ion initially positioned in the lateral loop region drifting into the solvent after a few hundred picoseconds at most.

In contrast to the parallel quadruplexes, the lateral T2A loops were very stable during the simulations. Loop 1 had a high initial rms deviation, but quickly stabilised to a conformation common to all simulations. The lateral T_2A loop residues formed a base triplet during the initial stages of each simulation. This triplet stacked with the G-quartets throughout the simulations.

A greater range of rms deviations was observed for the modified diagonal loops. Loop lengths of T_3 to T_5 gave stable conformations after initial deviations from the starting structure. The T_3 loop formed the same interactions as the T_2A diagonal loop in the NMR structure, with the first T being the furthest into solution, the second T stacking below on top of the terminal A, and the third T both stacking on the G-quartet and forming hydrogen bonds with the adenine (Figure 2.4.5).

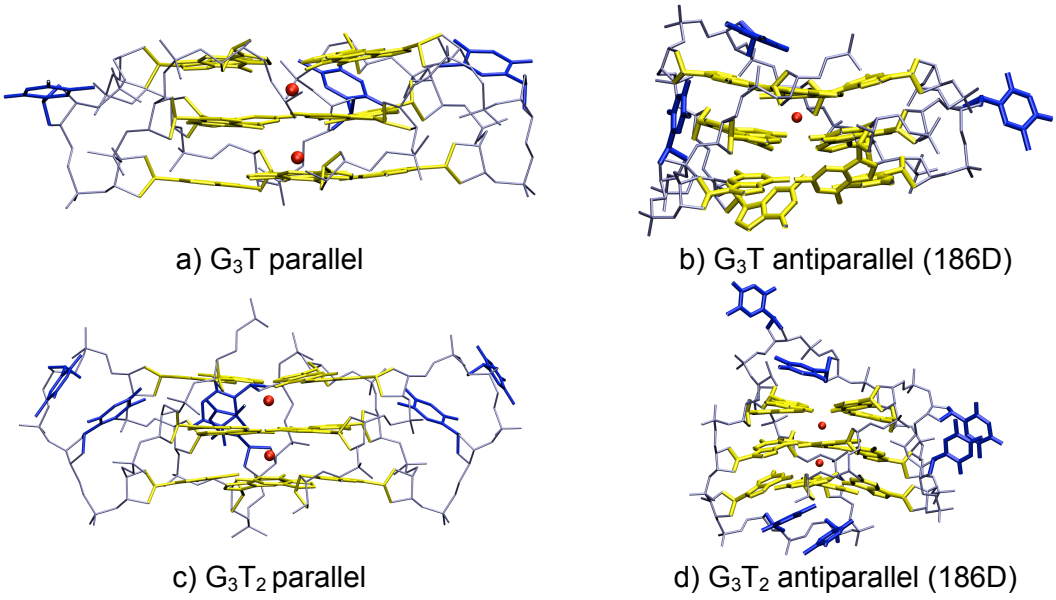
The first and last residues from the T_4 loop stacked on the G-quartets, with the second stacking on top of the triplet formed by the T and terminal A residues. The third T loop residue was stably located within the groove of the loop. The T_5 loop conformation is very similar, with one extra T stacking on top of the G-quartets. Only the T_6 loop had a higher flexibility. Stacking was still observed between the loop bases, but was more subject to fluctuations than in the shorter loop simulations.

Molecular dynamics simulations of multiple-loop variants. In order to further correlate the simulations with experimental data, multiple T_1 and T_2 loop variants were also simulated, and average structures are shown in Figure 2.4.6. The single and multiple loop variants differ in that TTA-G_3T is the only single loop variant quadruplex that shows a characteristically parallel CD spectrum, whereas both the G_3T and G_3T_2 multiple-loop variants appear to have some parallel character (with a peak around 265 nm).

MD simulations of parallel and antiparallel single and multiple-loop variants showed that only the G_3T sequence precluded an antiparallel fold. This simulation resulted in a large distortion of the G-quartets, with only 1 K^+ ion remaining within the core quadruplex channel as a result. It must also be born in mind that the antiparallel quadruplex selected for the simulations was based on the 186D PDB structure, which contains a mixture of parallel and antiparallel strands. The distortion caused to an antiparallel quadruplex with three lateral T loops would probably be much greater.

On the other hand, the TTA-G_3T antiparallel quadruplex was stable over the 4 ns simulation, confirming that a short lateral T loop is possible in an antiparallel

quadruplex, as long as there is only one present. Figure 2.4.6d shows that the antiparallel G_3T_2 quadruplex lower G-quartet was slightly distorted (curved), although to a much lesser extent than the antiparallel G_3T (Figure 2.4.6b). The hydrogen bonds between G residues in the G_3T_2 quadruplex were still maintained, and the channel K^+ ions were unaffected. This supports the proposal that the CD spectrum of the G_3T_2 quadruplex could be due to a mixture of both conformations.



a) G_3T parallel

b) G_3T antiparallel (186D)

c) G_3T_2 parallel

d) G_3T_2 antiparallel (186D)

Figure 2.4.6: Structures of the multiple-loop variants. Structures were averaged over the last 2ns of simulation. The colours are as in Figure 2.4.5.

Free energy calculations. The free energy differences between parallel and antiparallel quadruplexes of same loop lengths are shown in Table 2.4.4. The free energies were averaged over the last 3 ns of simulation. 4 ns was generally found to be a sufficient time scale to obtain stabilised enthalpic contributions to the free energy. The qualitative values obtained show that antiparallel quadruplexes are generally favoured, although the free energy differences are relatively small and often within error margins. The entropic contribution to the free energy differences was found to be non-negligible as it is almost equal to the enthalpic contribution.

	ΔH	TΔS	ΔG
TTA-G_3T	-12 (2)	+8 (1)	-4 (2)
TTA-G_3T_2 (143D)	+16 (2)	+11 (1)	+27 (2)
TTA-G_3T_2 (186D)	-11 (2)	+10 (1)	-1 (2)
TTA-G_3T_3	-14 (3)	+8 (1)	-6 (3)
TTA-G_3T_4	-19 (2)	+7 (1)	-12 (2)
TTA-G_3T_5	-12(2)	+7 (1)	-5 (2)
TTA-G_3T_6	-25 (2)	+14 (1)	-11 (2)
G_3T	-8 (2)	+6 (1)	-2 (2)
G_3T_2	-4 (2)	+6 (1)	+2 (2)

Table 2.4.4: Free energy differences in kcal/mol between antiparallel and parallel quadruplexes (antiparallel – parallel). Standard errors of the mean are indicated in brackets.

2.4.3 Discussion

Loop lengths play a major role in the stability of quadruplexes, as shown by the substantial decrease in melting points as loop lengths are increased for the multiple-loop variants. The G_3T sequence had the highest UV melting temperature. This is in contrast with melting studies carried out on d($G_4T_4G_4$) and d($G_4T_3G_4$) by Balagurumoorthy[64], in which the shorter loop length was found to destabilise the intermolecular hairpin dimeric quadruplex. The destabilising effect is much less apparent when only a single loop is varied, in which case all sequences have very similar melting points. This suggests that the TTA loops are able to compensate for one destabilising longer loop in the quadruplex. The resulting single-loop variant quadruplexes are however significantly destabilised relative to the native sequence.

Comparison of simulations and experimental results for T and T2 loop structures. For quadruplexes with short loop lengths, the simulations correlated well with experimental CD data. The results suggest that short single T loops can be tolerated in antiparallel quadruplexes, as long as there is only one, balanced

by other longer loops. The G_3T quadruplex is only able to fold in a parallel conformation, the antiparallel structure being distorted (Figure 2.4.6b), whereas the TTA-G_3T quadruplex can fold in both parallel and antiparallel conformations (with a lateral, rather than diagonal T loop). This was shown both in the simulations, and experimentally from the absence of peak around 295 nm in the G_3T CD spectrum.

On the other hand, the TTA-G_3T CD spectrum has peaks around both 260 and 295 nm, and simulations show that both parallel and antiparallel conformations are stable. A slightly longer loop length of 2 nucleotides enables both parallel and antiparallel conformations to form, whether one or all three loops are modified. The parallel TTA-G_3T_2 and G_3T_2 quadruplexes were both stable over the 4 ns simulations. Loop rms deviations were similar for both quadruplexes, however the G-quartets were slightly more stable in the G_3T_2 structure.

The antiparallel TTA-G_3T_2 and G_3T_2 quadruplexes were also stable during the simulations, as long as the short T_2 loops adopted a lateral, rather than diagonal geometry. However, the G_3T_2 average structure showed some distortion of the G-quartets (Figure 2.4.6d), although hydrogen bonding and stacking of the guanine bases was maintained. This was reflected in a higher rms deviation of the G-quartets for the antiparallel G_3T_2 structure compared to the antiparallel TTA-G_3T_2 structure (1.5 and 1.2 Å respectively).

Although shortening the loops leads to an increased stability (lower rms deviations) of the parallel structures (for both T and T_2 loops), the opposite is true for antiparallel loops. G_3T_2 can form an antiparallel quadruplex, but the simulations show that it is probably more likely to form a parallel structure. These results predict that structures such as that formed by part of the c-myc sequence, d($AG_3TG_3GAG_3TG_3$), with potentially two T loops and a GA loop, would be expected to favour a parallel conformation, as has been recently reported in revised topologies.[127,128]

Free energy calculations could not give much insight into the stability of the T and T_2 loop sequences. The parallel and antiparallel structures had very similar

free energies, as shown in Table 2.4.4. More unambiguously, however, the 143D antiparallel TTA-G_3T_2 quadruplex was largely disfavoured compared to the parallel (ΔG=+27 kcal.mol-1) and 186D antiparallel (ΔG=+28 kcal.mol-1) conformations. This was due to very high distortion of the G-quartets to accommodate the short diagonal T_2 loop. Such large free energy differences were not found for other short loop sequences. Both T and T_2 loops formed parallel and antiparallel quadruplexes of very similar free energies, indicating that all the structures could potentially form in solution.

Surprisingly, the antiparallel G_3T quadruplex was slightly more stable than the parallel structure, despite the distortion of one of the G-quartets, and only one K^+ remaining within the channel. This result does cast doubt about the ability of such free energy calculations to measure the relative stabilities of structurally diverse quadruplexes. The small ΔG values obtained could indicate that the structures are very close in energy, and are likely to be polymorphic in solution.

Comparison of simulations and experimental results for T3 to T6 loop structures. Longer loop lengths of both the single and multiple loop variants gave characteristic antiparallel conformation CD traces (Figures 2.4.2 and 2.4.3). However, these structures are more difficult to compare based only on molecular dynamics simulations, as both parallel and antiparallel conformations yielded stable trajectories. The loop residues in antiparallel structures did generally form more numerous and stable interactions between themselves and with the G-quartets. Due to the geometry of strand-reversal loops, interactions between the loop residues and the G-quartets are more difficult to form.

Free energy calculations did show a preference for antiparallel over parallel conformations, consistent with the CD spectroscopic analysis, although the small free energy differences must be interpreted with caution (Table 6). The antiparallel quadruplexes were found to be between 1 and 12 kcal.mol^{-1} more favourable than the parallel quadruplexes. Small values indicate that both conformations could coexist in solution, even in the longer loop length cases. This is not observed in CD spectra, where for loop lengths of 3 or more, no characteristic parallel quadruplex peaks are observed.

Importance of interactions within the loops. In more general terms, the MM-PBSA calculations showed that the enthalpic contribution favours antiparallel quadruplex structures, whereas the entropic contribution favours parallel structures. The antiparallel structures formed a greater number of stable loop-loop and loop-tetrad interactions, which could explain the more favourable enthalpic contribution. In all three loops, the bases could stack with each other and with the G bases, and long-lived hydrogen bonds were formed between loop residues. Stacking interactions between loop residues in the parallel quadruplexes were generally observed, although the residues involved in these interactions varied during the simulations, and individual interactions were relatively short-lived.

The multiple-loop variants had smaller ΔH values, indicating that there were fewer differences in loop interactions between the parallel and antiparallel conformations. Smaller loops are therefore not as enthalpically favoured in antiparallel quadruplexes as longer loops are. On the other hand, entropic factors are more significant in parallel structures, in which loop mobility is greater compared to antiparallel structures. Generally, the rms deviations of loops with respect to the average structure are larger for parallel than antiparallel quadruplexes.

Interactions within the loops probably play a major part in the preferential antiparallel folding of quadruplexes with loop lengths of three or more. In the 143D antiparallel quadruplexes, it was shown that the two T_2A lateral loops can form a triplet which stacks with the G-quartets, in addition to stacking of the diagonal loop residues. Fewer stabilising interactions are possible within the parallel conformation. When the loop residues are removed and replaced by non-nucleosidic linkers, as shown by Risitano and Fox,[135] CD spectra of unimolecular quadruplexes exhibit characteristic parallel conformations peaks. However, the use of base-containing loops in the present work appears to favour the formation of antiparallel structures.

If several quadruplex structures do form in solution for a given sequence, the antiparallel quadruplex is probably the thermodynamically favoured conformation, especially for longer loops. Risitano and Fox also showed that a quadruplex sequence with the G_3T repeat was too stable for the melting point to be measured, and that a quadruplex with G_3T_2 repeats had a biphasic curve in K^+ solution.[135] This agrees with our simulations in which G_3T is only stable in a parallel conformation, whereas G_3T_2 can fold in both parallel and antiparallel conformations.

Phan and Patel[25] showed that both parallel and antiparallel conformations of a related sequence, d(TAGGGTTAGGGT), have different rates of folding and unfolding, but both coexist in K^+ solution. Risitano and Fox[135] have observed such parallel quadruplexes even when there are no loop residues. The interactions between loop residues and G-quartets might be one of the factors that drive the equilibrium towards formation of antiparallel structures. This could explain the dominance of antiparallel conformations in the CD spectra, especially as the loop lengths increase. There is growing evidence that the loops contribute an important amount to G-quadruplex stability, and comparing these results with the behaviour of non-residue-containing loops[135] shows that the loops and not the G-quartets determine the dominant fold of particular sequences.

Due to the structural polymorphism of G-quadruplexes, it was beyond the scope of this study to simulate all possible antiparallel structures that could be adopted by these quadruplex-forming sequences. The two templates were chosen due to the availability of experimental structures that did not require changes to the G-quartets. Simulations tend to be very dependent on the starting structures used, and changes were therefore limited to the loops. For each loop length we endeavoured to find a stable conformation, and not to discriminate between all the possible loop structures. The free energy calculations were carried out for a single antiparallel structure, although of course other potentially more favourable conformations cannot be ruled out.

Recently, there have been some questions as to the ability of current force fields to represent G-quadruplex loops accurately.[153] The present work did not aim to

find the “best” loop geometries for each of the quadruplexes studied, but only to show that a certain quadruplex could potentially form. Stable loop conformations were not obtained for several of the longer loop variants in this study, however the simulations showed that the G-quartets were stable despite this. The distortions of certain short loop G-quartets are expected to reflect genuine inabilities of these folds to form, rather than shortcomings of the force field.

2.4.4 Conclusions

We have shown that a combination of experimental and molecular modelling methods can yield valuable information about the structure of G-quadruplexes with varying loop lengths. CD spectra suggest that parallel quadruplexes are only present when the loops are too short to accommodate an antiparallel conformation. G_3T was the only sequence studied which could only exist as a parallel quadruplex. When only one single-T loop was present, both parallel and antiparallel quadruplexes were stable. Simulations showed that parallel conformations are stable for all of the sequences studied, and are likely to be very close in energy to antiparallel conformations. Assuming that parallel structures are present in solution only when characteristic peaks are observed in CD spectra could be misleading. The following rules for quadruplex folding can be drawn up following the results obtained from simulations and experimental data:

- Multiple single-base loops can only form parallel structures
- Multiple two-base loops can form both parallel and antiparallel structures, however parallel conformations are likely to be favoured
- An individual single-base loop enables the formation of both parallel and antiparallel structures, but the parallel structure is more favourable
- Longer loops can also exist as parallel or antiparallel conformations, however the antiparallel structure is likely to be preferred.

Melting studies showed that shorter loop lengths are energetically favoured, with G_3T_2 having the highest measurable melting point. The free energy calculations are only relative values, and cannot be used currently to compare quadruplexes

of different loop lengths to assess their stability relative to the number of loop residues. However, the simulations do show that longer loops have higher rms deviations, and fewer long-lived interactions.

2.5 Folding rule and quadruplex survey

By examining all of the results and considerations described in the previous sections, it was possible to develop and propose a 'folding rule' for quadruplexes. This is as follows:

> 'Any sequence of the form d(GGG…GGG…GGG…GGG),
> where … represents a gap of between 1 and 7 bases (possibly including G),
> and GGG represents a series of at least 3 G residues,
> will form a quadruplex structure under appropriate conditions'

The 'appropriate conditions' referred to above are the experimental conditions used throughout, namely with the DNA in dilute (1-5 μM) solution at room temperature in GQ buffer (100 mM KCl and 10 mM Tris.HCl pH 7.4). Evidence for quadruplex formation could be an appropriate CD trace, UV melting transition at 295 nm, or any other appropriate method such as NMR spectroscopy.

Equipped with this rule, it was possible to examine the sequences of a sample of genes of biological interest, to investigate whether they would have putative quadruplex sequences (PQSs) in or near the genes. The regions studied include all introns and exons, and 1000 bp up-and down-stream of the gene as labelled in ENSEMBL. On the basis of their relevance for cancer, the sequences associated with the genes in table 2.5.1 below were analysed in more detail. A complete study of the whole genome is described in Chapter 3.

A considerable number of putative quadruplex-forming sequences in these genes were found, every single one of these genes having potentially relevant sequences. They are listed by gene below in table 2.5.2 and classified according to their location and strand (Template/antisense strand or coding/sense strand).

The sequence shown in each case is that for the coding/sense strand; hence, those sequences located in the template/antisense strand are shown as C-rich sequences complementary to the actual strand.

Name	Description
C-myc	Participates in the regulation of gene transcription. Activates the transcription of growth-related genes. Overexpression of MYC is implicated a variety of hematopoietic tumors and leukaemia.
H-ras	Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Defects are a cause of bladder cancer and oral carcinomas
RhoB	Regulates a signal transduction pathway linking plasma membrane receptors to the assembly of focal adhesions and actin stress fibers.
N-ras	Ras proteins bind GDP/GTP and possess intrinsic GTPase activity. Defects in N-ras are a cause of juvenile myelomonocytic leukaemia.
c-kit	Receptor for stem cell factor (mast cell growth factor). It has a tyrosine-protein kinase activity. Defects in c-kit are a cause of piebaldism and gastrointestinal stromal tumors.
P53	Acts as a tumor suppressor in many tumor types. P53 is mutated or inactivated in about 60% of cancers
Telomerase	Telomerase is a ribonucleoprotein reverse transcriptase essential for the replication of chromosome termini in most eukaryotes. It elongates telomeres. Activation of telomerase has been implicated in cell immortalization and cancer cell pathogenesis.

Table 2.5.1. List of genes manually examined for PQS presence. A standard abbreviation for each is shown, followed by a brief description of their physiological role. Information is taken from the SWISS-PROT database.

Location	Strand	Sequence
C-myc		
5' upstream	Templ.	CCCCACCTTCCCCACCCTCCCCACCCTCCCC
5' UTR	Templ.	GGCCCAGCCCTCCCGCTGATCCCCCA
Exon	Templ.	GCTGCCCACCCCGCCCCTGTCCCCTAG
Intron	Coding	ACGGGGGCGGTACTGGGGGTGGGGACGGGGGCGGT
Intron	Templ.	CCTCCATCCCTTCCCCTTAGACTGCCCATG
Intron	Templ.	CAGCCCCCCTCCCCGTTTGTCTCCCACCCCTCAGG
H-ras		
5' upstream	Coding	TTTGGGGGGACTGGGCGGGCAGGG
5' upstream	Coding	GGGGTGGGTTTGCGGTGGGAGTAGGGGAGCTGGGG
5' upstream	Templ.	CCCGCCTTGGTCCCGGCGGATCCCAGCCTTTCCCCAGCCCGTAGCCCC
5' upstream/ 5' UTR	Templ.	CCGCCCGTGCCCTGCGCCCGCAACCCGAG
5' UTR	Templ.	GCGCCCCGCCCCCGCCCCGCCCCGGC
Intron	Coding	CGCGGGCCGGGGGCGCGGGGCCGGCGGGCGT
Intron	Coding	CGGGTGGGTGGGGCCGGGCGGGGCCC
Intron	Coding	TGTGGGGCCTGGGCTGGGCCTGGGCCT
Intron	Templ.	GGACCCGCCCCGCCCCGCCCCAGG
Intron	Templ.	GGACCCCCCGGGACCCATGTGACCCAGCGGCCCCTCG
Intron	Coding	CCCGGGACGGCAGGGCAGTGAGGGAGGCGAGGCGGGGGCCGGGGTATGGGCTC
Intron	Templ.	TGACCCTGGGGCCCGGCCCCTCTTGTCCCCAGA
Exon	Templ.	GAACCCCAGCCCTTAGCTCCCCTCCCAGG
RhoB		
5' upstream	Templ.	CTGCCCCTCCCCCACAGGCCCGACCACCCGCC
N-Ras		
5'-UTR	Coding	TGTGGGAGGGGCGGGTCTGGGTGC
Intron	Templ.	CTTCCCTCCCTCCCTGCCCCCTTACCCTCC
Intron	Coding	TTTGGGGGGGGTTGGGGGGGATGGAG

c-kit		
5'-upstream	Coding	CCCGGGCGGGCGCGAGGGAGGGGAGG
5'-upstream	Coding	AGAGGGAGGGCGCTGGGAGGAGGGGCTG
Intron	Coding	TCAGGGCTGCCTGGGGGTGGGGTGGGAGA
Intron	Templ.	AGACCCCTCCATCCCTCCCGCCCCCATCCCCACT
Intron	Templ.	CTTCCCTCCCTCCCTCCCTCCCTCCCTCCCCCAATCCCTGT
Intron	Coding	AGCGGGGATGGGGTGGGGGGTGCGGGGAAACGGGGGACT
Intron	Coding	TGTGGGTGGGGTGGGGTGGGGTGGGGTGG
Intron	Coding	CCAGGGTGGTGGGGGGAGGGCGGAGGCGGGAGG
P53		
Intron	Coding	GCTGGGGCTGGGGGTGGGGCAGTGGGGACT
Intron	Coding	GGTGGGGAAGGGTGGGGGCCAAGGGGGTGT
Intron	Templ.	AGGCCCACCACCCCCACCCCAACCCCAGCCCCCTAG
Intron	Coding	GGAGGGCTGGGGACCTGGAGGGCTGGGGGGCTGGGGGGCTG
Intron	Coding	ACTGGGGTCTCTGGGAGGAGGGGTTAAGGGTGG
3' UTR	Templ.	CCTCCCACCCCCATCTCTCCCTCCCCTGC
3' UTR	Templ.	TCACCCCACCCTTCCCCTCCTTCTCCCTTT
Telomerase		
5' upstream	Templ.	GGACCCCGCCCCGTCCCGACCCCTCCCGGGTCCCCGGCCCAC CCCCTCCGGGCCCTCCCAGCCCCTCCCCTTC

Table 2.5.2. PQS found by manual searching in various genes of interest. In each case, the sequence was searched from 1000 bp upstream of the trancription start site until 1000 bp of the transcription stop site, except for telomerase, which due to its length was only examined as far as the second intron. The first column describes the genomic location of the quadruplex, and the second column describes the strand that can form a quadruplex (template or coding). The third column shows the sequence, in all cases being given for the coding strand, which means that where there is a quadruplex-forming sequence in the template strand, it is shown as C-rich.

An example of the layout of these sequences around the 5' end of one of these genes, Ha-ras, is shown in figure 2.5.1. A number of potential roles could be proposed for each of these sequences:

- They could be involved in downregulation of gene transcription as proposed for c-myc, such as by acting as a steric block to transcription initiation
- They could upregulate gene transcription by recruiting transcription factors
- They could serve to delineate the transitions between intron and exon, or 5'UTR and exon.
- They could be involved in splicing

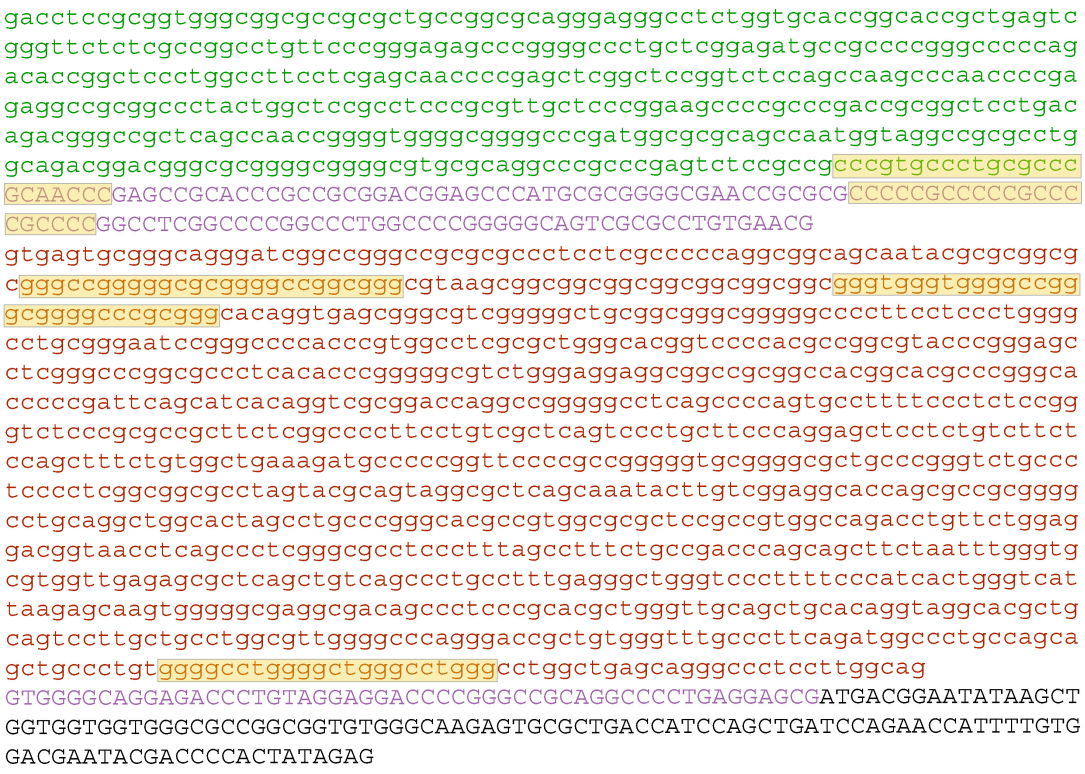


Figure 2.5.1 Example data from ENSEMBL. Data is shown for the 5' end of the Ha-ras gene. In green is shown the 5' upstream region. The purple region is the 5' UTR, which is transcribed but not translated. The black letters are the coding region, beginning with an ATG start codon (AUG as DNA). In red is an intron. Quadruplexes are shown inside yellow boxes.

2.6 Biophysical studies on putative quadruplexes

2.6.1 Sequence selection

As most of these genes are overexpressed in cancer, a few sample sequences were selected that were thought to be the most likely candidate targets for small-molecule mediated downregulation of the gene. These sequences are shown in table 2.6.1, and were studied to confirm that they did form quadruplexes. C-myc was included in the list as a positive control, in view of the work by Hurley and

co-workers.[59] A form of the human telomeric quadruplex, $d(GGTTAG)_5$ (**hTelo**) was also included as a positive control, and a mutated form $d(AGTTAG)_5$ (**aTelo**), in which the central guanine of each tetrad had been mutated to an adenine such that it cannot form a quadruplex, was used as a negative control. DMS footprinting was used to confirm that **aTelo** did not form a quadruplex (see figure 2.6.1).

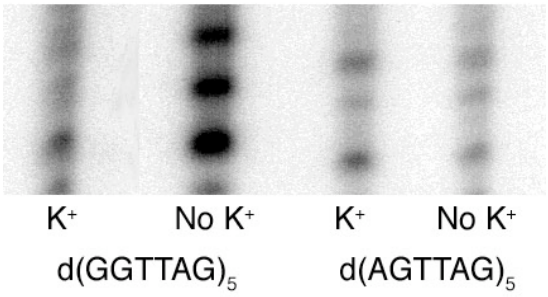


Figure 2.6.1. DMS footprinting studies on $d(GGTTAG)_5$ (**hTelo)** and $d(AGTTAG)_5$ (**aTelo)**. Oligonucleotides were phosphorylated with ^{32}P, treated with DMS for 1 minute, and then soaked in piperidine, and then analysed using PAGE. The **hTelo** sample shows a clear difference in cleavage intensity upon the addition of potassium before DMS treatment, corresponding to protection of the guanine N_7 position that is otherwise methylated. **aTelo** shows no difference with or without potassium.

Name	Sequence
Ha-ras	TGGTGGGCCTGGGGCTGGGCCTGGGCT
N-ras	TGTGGGAGGGGCGGGTCTGGGTGC
c-kit-1	AGAGGGAGGGCGCTGGGAGGAGGGGCTG
c-kit-2	CCCGGGCGGGCGCGAGGGAGGGGAGG
c-myc-1	TGGGGAGGGTGGGGAGGGTGGGGAAGGTGGGGAGG
c-myc-2	TGGGGAGGGTGGGGAGGGTGGGGAAGG
hTelo	GGTTAGGGTTAGGGTTAGGGTTAGGGTTAG
aTelo	AGTTAGAGTTAGAGTTAGAGTTAGAGTTAG

Table 2.6.1. Sequence of putative quadruplexes selected for biophysical studies

BLAST analysis was used to investigate how frequent each of these motifs was in the human genome. **C-kit-1** and **C-kit-2** were both unique; **Ha-ras** occurs one other time, **N-ras** three other times and **C-myc** a total of five times.

2.6.2 Biophysics

All of these sequences were studied by UV melting and circular dichroism spectroscopy. Before examination, the quadruplexes were allowed to form in equilibrium by heating 4 μM samples to 90°C and then allowing them to slow-cool at a constant rate of 0.1 °C/min to 5°C, using a modified heat block. It was found that in some cases rapid cooling prevented quadruplex formation from occurring.

All of the sequences except **aTelo** exhibited the classic hypochromic transition at 295 nm that is a hallmark of quadruplex melting.[129] The melting temperatures were very variable, and in the case of **c-kit-2** and **c-myc-2**, the melting transitions were so high that they were not resolved experimentally. To confirm that this was due to the formation of a highly stable quadruplex rather than the absence of quadruplex structure, the experiments were repeated at lower (20 mM) K^+ concentrations to reduce the stability of the quadruplexes and hence the melting point.[129] This did reduce the melting temperature to manageable levels. The melting temperatures were obtained using van't Hoff analysis and by the derivative method, and are shown below.

Name	[K^+] in buffer (/mM)	T_m (/°C)
Ha-ras	100	67
N-ras	100	66
c-kit-1	100	52
c-kit-2	100	Not observed
	20	58
c-myc-1	100	56
c-myc-2	100	Not resolved
	20	76
hTelo	100	53
aTelo	100	Not observed

Table 2.6.2. Melting temperatures for putative quadruplexes. Annealed samples at 5 μM concentration were slowly heated from 10 °C to 90 °C and cooled back to 10 °C. The absorption at 295 nm showed a hyperchromic transition, which was analysed by the van't Hoff method. Errors in T_ms are estimated at ± 2 °C.

Circular dichroism measurements were also performed on all of these samples. When the slow annealing protocol was not performed, none of them except **hTelo** showed the characteristic peaks of one of the two forms of quadruplex. This suggests that the kinetics of folding are relatively slow, or that there may be 'misfolded' states that are populated with relatively rapid cooling. However, when the slow annealing protocol was used, they all showed one of the characteristic CD traces for a quadruplex,[64,87,139] with the exception of **aTelo**, which had the trace expected for a G-rich unstructured single-strand of DNA. [139]

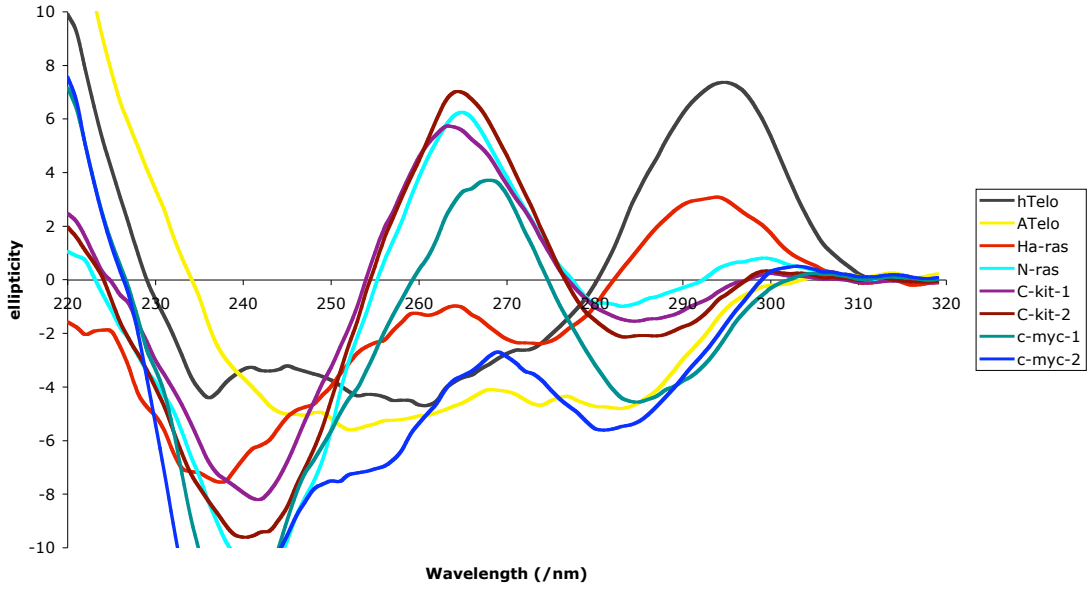


Figure 2.6.2. Circular dichroism spectra of the putative quadruplex sequences. **aTelo** (yellow) exhibits no secondary structure. **hTelo** (black) shows a peak at 295 nm, corresponding to an antiparallel quadruplex. **N-ras, c-kit-1, c-kit-2 and c-myc-1** (cyan, purple, brown and green respectively) show peaks around 265 nm, corresponding to parallel structures. **Ha-ras** (red) shows a mixture of forms, with peaks around 265 and 290 nm. **C-myc-2** shows a weak parallel signal, suggesting it may only be partially formed.

As discussed earlier in this chapter, sequences with short (especially single-base) loops are predicted to preferentially form parallel quadruplexes, and this is seen here. **N-ras, c-kit-1, c-kit-2, c-myc-1 and c-myc-2** all contain single-base loops, and are found to form parallel sequences. **hTelo**, with three-base loops, forms an antiparallel sequence according to the CD data, and **Ha-ras**, with a 2-base loop and two three-base loops, appears to be a polymorphic mixture of the two forms. This is encouraging evidence for the results earlier in this chapter, and shows they may be applicable even with non-thymine loops.

2.6.3 N-ras – a case example

A more detailed series of studies into the behaviour of the **N-ras** sequence has also been carried out. These studies involved varying the DNA concentration, the nature of the cation used, and the salt concentration. A series of point mutations in the **N-ras** sequence was also performed. The experiments were all performed by Shuizi Yu, a very able part III student I supervised, and analysis (following the Van't Hoff procedures described in chapter 5) was performed by her, all under my supervision.

2.6.3.1 DNA concentration dependence

There are two principle thermodynamic methods for determining whether a quadruplex is intramolecular or intermolecular. Both rely on the manner in which the equilibrium constant changes with different molecularity. In the analysis of a unimolecular sequence, the expression for the equilibrium constant, K_{eq}, in terms of the fraction of DNA folded, α, is given by equation 1 below. For a bimolecular system, in which two single-stranded units combine to form a quadruplex, the equivalent expression is shown by equation 2 below, which introduces a third term, C_T, the total concentration of DNA present. Similar expressions may be derived for a tetramolecular system.[154]

$$K_{eq} = \frac{\alpha}{(1-\alpha)} \qquad \text{Equation 1}$$

$$K_{eq} = \frac{\alpha}{2C_T(1-\alpha)^2} \qquad \text{Equation 2}$$

These expressions may be used to differentiate between different molecularities in two ways. Firstly, when fitting an individual UV melting curve using the Van't Hoff method, either relationship may be used, and the quality of the fit compared. A more detailed approach is to vary the total DNA concentration, C_T, and perform a series of UV melting experiments. For a unimolecular system, as there is no dependence on C_T, the melting temperatures should be unaffected by changing concentration (at relatively low concentrations – at high concentrations there may be a transition from unimolecularity to oligomolecularity). However, a

bimolecular (or higher order) system should exhibit a clear increase of melting temperature with concentration.

In order to confirm the molecularity of **N-ras**, which had previously been concluded based on the Van't Hoff method, we performed a series of experiments in which the concentration of ssDNA was varied from 1 to 10 μM. Thermodynamic analysis was then performed as described previously.

DNA concentration (/μM)	T_m (/°C)
1	65
2	66
4	65
10	65

Table 2.6.3 Melting temperatures for **N-ras** at various concentrations, measured from UV thermal spectroscopy. Experimental error is estimated at ±1 °C for all UV melts.

This shows no dependence on the concentration, supporting the conclusion that this sequence forms an intramolecular quadruplex, at least in the concentration range studied.

2.6.3.2 Varying the salt and salt concentration

Quadruplex stability varies considerably with the concentration of potassium[155], and thermodynamic analysis of this allows the estimation of the number of potassium ions released (or bound) during the transition quadruplex/ssDNA transition.[155] This arises because the equilibrium shown below results in a relationship between K and ΔG which has a dependence on Δn, the change in potassium binding (Equation 3).

$$\text{Quadruplex} \rightleftharpoons \text{ssDNA} + \Delta n\ K^+$$

$$\frac{d\Delta G}{d\ln[K^+]} = -\frac{\Delta n}{RT} \qquad \text{Equation 3}$$

A series of experiments was performed in which the concentration of potassium ions was varied from 0 to 200 mM, and the UV thermal melting transitions were analysed; the melting temperatures are shown in table 2.6.4. Plotting T_m or ΔG against ln[K^+] gave a straight line, as seen elsewhere.[155] Analysis of the gradient of ΔG against ln [K^+] gave a value for Δn of +4.0. This implies that there are four K^+ ions bound in the quadruplex form that are released upon melting. It cannot be concluded with any certainty where these might be in the absence of a crystal structure, but it may be hypothesised that each of the three tetrads has a K^+ ion above and below the plane, as seen in the crystal structure of the human telomeric sequence, solved by Neidle and coworkers.[29] An alternative hypothesis might involve some of them being bound by the loops. [23,155]

K^+ concentration (/mM)	**T_m (/°C)**
0	N/A
10	43
50	63
100	66
200	75

Table 2.6.4 Melting temperatures for **N-ras** at various KCl concentrations, measured from UV thermal spectroscopy. No melting transition was observed without KCl. T_m values have an estimated error of ±1 °C

In contrast to duplex DNA, where the ionic strength is of key importance, rather than the identity of the ion, quadruplexes are highly specific as to the cations they prefer. A number of studies have been performed, and in general quadruplex sequences appear to form in the presence of monovalent cations, with decreasing stability $K^+ > Na^+ >> Li^+$.[65,143] This has been shown to correspond with the specificity of alkali metals binding the simple guanine monophosphate, and is thought to reflect cavity binding in the tetrads.[156] K^+ appears to be the optimal size for binding.

In order to confirm this applied to the structure adopted by N-ras, another series of experiments was performed in which different cations were used, in each case

as 100 mM solutions of the chloride salt. Both **N-ras** and **hTelo** were studied for comparison, and the results are shown in table 2.6.5.

Cation (100 mM)	N-ras T_m (/°C)	hTelo T_m(/°C)
K^+	66	63
Na^+	34	56
Li^+	N/A	37
NH_4^+	N/A	N/A

Table 2.6.5 Melting temperatures for **N-ras** and **hTelo**, using UV spectroscopy. In each case, the buffer contained 10 mM Tris.HCl pH 7.4 and 100 mM XCl, where X is a monovalent cation. The entries marked N/A refer to systems in which no melting was observed. Errors in T_m values are estimated as ±1 °C.

These results confirm the traits previously described for other quadruplex sequences. One interesting feature to note is that although in the presence of 100 mM KCl **N-ras** and **hTelo** have very similar melting temperatures, the stability of **N-ras** decreases more when the cation is changed. Since the core structure, where the cations are presumed to bind, has the same sequence in each case, the reason for the difference must lie either in the conformation of the tetrads or in binding sites interacting with the loops. Further detailed studies of other sequences would be necessary to fully rationalise this pattern, but one possible theory is that there may be a difference between parallel and antiparallel sequences as to the dependence on the nature of the cation. This could arise because of the differing geometrical arrangement of the cavities formed by parallel strands, with guanines all *anti*, and by antiparallel strands, with guanines *syn.syn.anti.anti*.

2.6.3.3 Mutant studies

In order to further investigate the properties of the **N-ras** sequence, a series of mutants was studied, each intended to test particular aspects of the structure – the effect of runs of guanine greater than three and of varying loop lengths. These are listed in table 2.6.6, along with the melting temperatures produced from UV spectroscopy, and the difference in T_m from the parent **N-ras** sequence.

Name	Sequence	T_m (/°C)	ΔT_m (/°C)
N-ras	TGTGGGAGGGGCGGGTCTGGGTGC	66	-
G8A	TGTGGGA**A**GGGCGGGTCTGGGTGC	61	-5
G11A	TGTGGGAGGG**A**CGGGTCTGGGTGC	63	-3
G8A;G11A	TGTGGGA**A**GG**A**CGGGTCTGGGTGC	38	-28
G15A	TGTGGGAGGGGCGG**A**TCTGGGTGC	N/A	N/A
T18G	TGTGGGAGGGGCGGGTC**G**GGGTGC	72	+6
Ins7T;Ins12T	TGTGGGA**T**GGGGC**T**GGGTCTGGGTGC	56	-10

Table 2.6.6; Sequences studied as mutations of the **N-ras** parent sequence, along with melting temperatures and the difference from the parent sequence. **G15A** did not show a melting transition. All T_m values have an estimated error of ±1 °C.

The **N-ras** sequence consists of four sets of guanine runs, all but the second having precisely three guanines; the second run having four. The gaps between the guanine repeats are, in the order 5'-3', 1,1,3. Interestingly, the two mononucleotide gaps occur either side of the G_4 unit, and in practice one of them may well be a dinucleotide loop incorporating one of the guanines, as shown for **c-myc**[127,128]. Hence there may be two potential forms of the sequence in solution (ignoring parallel/antiparallel structure), which may be written as:

1) TGTGGGAGGGGCGGGTCTGGGTGC Loops 2,1,3

2) TGTGGGAGGGGCGGGTCTGGGTGC Loops 1,2,3

A series of circular dichroism spectra were also collected for each mutant oligonucleotide, with samples prepared by slow cooling at 4 mM in GQ buffer. Spectra were collected at 4 °C. The results are shown in figure 2.6.3. The parent **Nras** sequence exhibits a positive peak at 260 nm and a trough at 240 nm, as does the **T18G** mutant. This is consistent with the formation of a parallel-stranded quadruplex.[64] The dual mutant **G8A;G11A** exhibits a peak at 295 nm, consistent with the formation of an antiparallel quadruplex.[64] The other mutants studied appeared to exhibit a superposition of both spectra, suggesting that both structures existed in equilibrium.

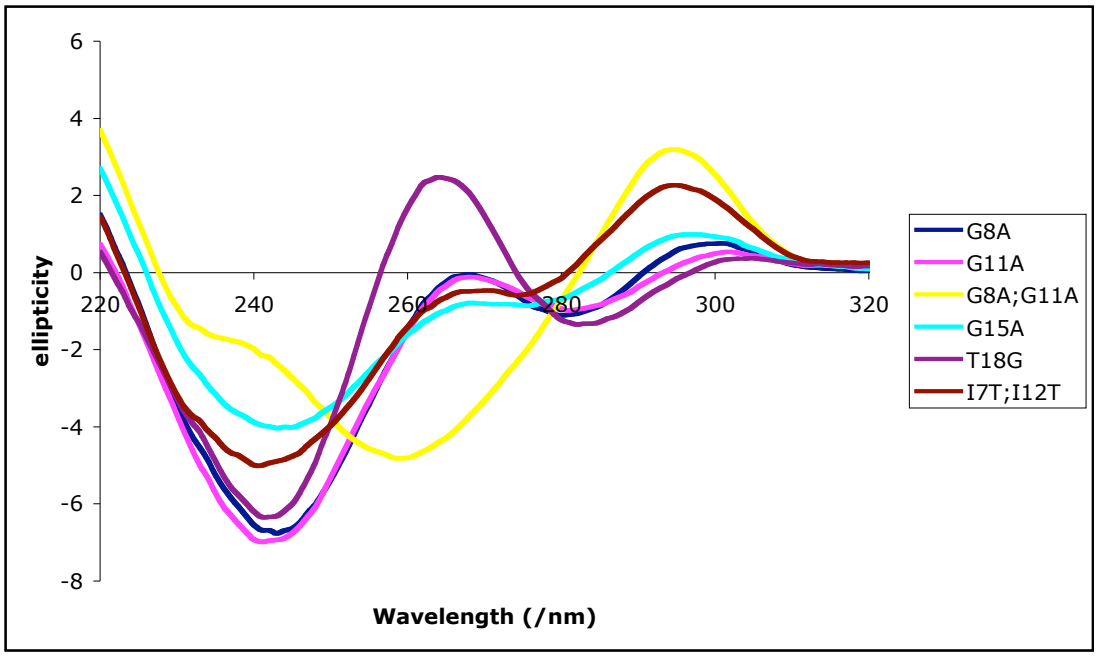


Figure 2.6.3. CD spectra for the **Nras** mutant oligonucleotides. Data was collected at 4 μM concentration at 4 °C. The **T18G** mutant exhibits a positive peak at 260 nm and a trough at 240 nm. The dual mutant **G8A;G11A** exhibits a peak at 295 nm. The other mutants studied exhibit a superposition of both spectra.

The first three mutants, **G8A, G11A** and the double mutant **G8A;G11A**, were designed to test which of these arrangements were more plausible, or whether the structure was composed of a superposition of the two states. The UV melting results show that mutation of either of the terminal guanines has a moderate effect on the overall stability (ΔT_m –5 °C and –3 °C), but that mutation of both simultaneously has a significant effect (ΔT_m –28 °C). This suggests that both structures can exist, and have relatively similar stabilities. They may be in dynamic equilibrium, depending on the kinetics of the system. This suggests a scenario similar to that found by both Hurley and Patel for **c-myc**.[127,128]

The mutant **G15A** demonstrates the significance of having at least three guanines in a row to form stable quadruplexes. Interestingly, this mutation seems to have a more significant effect (No melt observed) than the **G8A;G11A** double mutant (ΔT_m –28 °C), although they both remove one set of three guanines.

The fifth mutant, **T18G**, further tests the hypothesis that sequences with four consecutive guanines in one of the guanine tracts are more stable than those

with only three, because they can adopt either of two structures. An increase in T_m of six degrees is a significant degree of stabilisation, and comparable to the loss of stabilisation in **G8A** and **G11A** (ΔT_m –5 °C and –3 °C respectively).

The last mutant, the dual insertion mutant **Ins7T, Ins12T**, explores the effect of lengthening the loops linking the guanine runs. In the study of oligo-dT sequences described in §2.4, it was shown that longer loops result in less stable sequences, and this is confirmed here, with the elongation of the first and second loops by one base each resulting in a ΔT_m of –10 °C.

2.6.4 Ligand binding

Within our group and elsewhere, we have developed a wide selection of small molecules and proteins, all of which were designed and selected to bind the quadruplex formed by the human telomeric repeat, $d(GGGTTA)_n$. Another form of evidence to show that other genomic sequences also form quadruplexes is to demonstrate that they are recognised by such quadruplex ligands.

Many of the small molecules that bind quadruplexes are fluorescent. This means that changes in their fluorescent properties can be used as a measure of the amount of binding. There are two principle properties that could be used. The fluorescent brightness (quantum yield) of a dye varies with the environment it is in. Solvent exposure tends to decrease the brightness, π-systems tend to also decrease it. In general, it is hard to work out *ab initio* what the sign of a binding change will be, but where there is one, it acts as a useful reporter.

Another potential property is the fluorescence anisotropy.[157] This refers to the effect the dye has on plane polarised light. A high value refers to a highly anisotropic species – in other words, one that is rotating slowly. Low values refer to more isotropic species, which rotate rapidly. Since the rate of rotation of a dye bound to a quadruplex must be slower than that for free dye, this can also be a useful reporter. However, because anisotropy measurements are noisier that fluorescence measurements, the former approach is preferred where possible.

Our group has developed a class of cyanine dyes that bind to the human telomeric quadruplex. My student, Shuizi Yu, has performed titration experiments in which a concentrated solution (500 μM) of pre-annealed **N-ras** was titrated into a solution of the dye Cy3 1-ethylbenzoxazole against 1,3,3-trimethylindole (see figure 2.6.4) at a concentration of 1 μM. The fluorescence output was monitored, and shown in figure 2.6.5. It shows a significant increase in fluorescence upon binding, and was fit to a simple binding rule equation, to yield a binding constant K_d= 33 ± 5 μM.

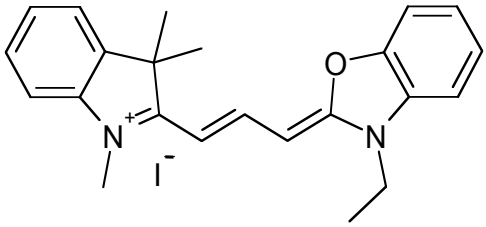


Figure 2.6.4. A Cy3 dye with heterocycles 1-ethylbenzoxazole and 1,3,3-trimethylindole. Synthesised and donated by Jen Hake.

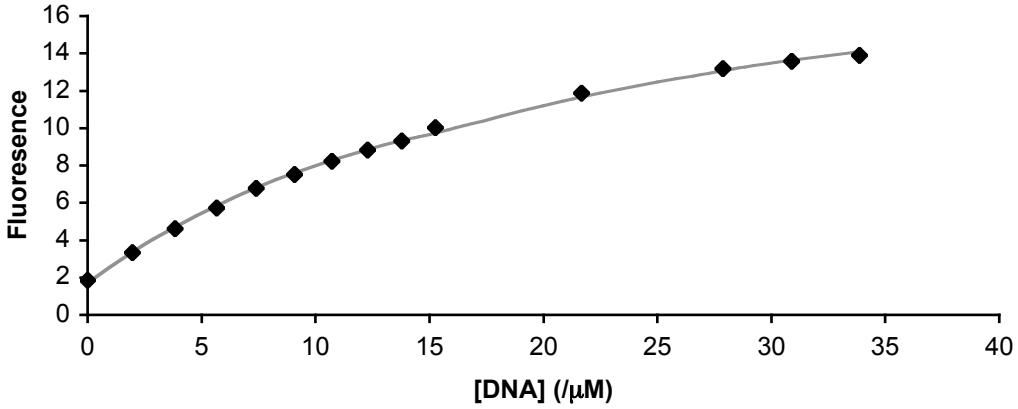


Figure 2.6.5. Fluorescence binding curve for the Cy3 dye against **N-ras**. A pre-annealed solution of DNA (100 μM initially, then 400 μM for the last four additions) was added to 100 μl of a fixed amount of dye (1 μM), and the fluorescence measured using an Aminco-Bowman spectrophotometer, summing between 520 nm and 550 nm. Black diamonds are the measured data. The grey line is a fit to a simple binding rule, and gives a binding constant of around 33 μM.

The fact that the fluorescence increases significantly on binding, from 2 to 34 units (extrapolating to complete binding), suggests that the dye molecule binds in a groove or loop of the target quadruplex, shielding it from solvent quenching but not quenching it with the π-rich guanine bases. However, it is extremely difficult

to make any clear conclusions from these experiments as to the binding mode, and any such conclusion should be treated as highly tenuous.

Other members of our group have studied the binding of various small molecules and the zinc-finger protein Gq1[92] to these novel quadruplex sequences, using the techniques of Surface Plasmon Resonance and Enzyme-Linked Immunosorbent Assay (ELISA).The SPR results, performed by Dr. Sylvain Ladame, showed that some small-molecule ligands were capable of binding to **c-kit-1** with binding constants similar to those observed to **hTelo**. The ELISA studies were performed by Dr James Redman, and the binding constants obtained are shown in table 2.6.3. These show that the engineered zinc finger protein Gq1 binds the novel quadruplexes tightly, and does not bind single-stranded **aTelo** or the duplex **Zif** sequence, which the prototype zinc finger protein Zif-268 bound specifically.

Sequence	**Htelo**	**hTeloGG**	**c-kit-1**	**c-kit-2**	**Ha-ras**	**N-ras**	**aTelo**	**Zif**
K_d (/nM)	71	56	11	11	18	18	-	-

Table 2.6.3. Results of ELISA using Gq1 against various pre-annealed DNA sequences. K_ds were calculated using a Langmuir model. **hTelo** and **hTeloGG** differ as rotational isomers of the human sequence $(GGGTTA)_n$. **Zif** refers to the native substrate for Zif-268, the transcription factor from which Gq1 was derived. Neither it nor **aTelo** showed any binding. Performed by Dr James Redman.

These observations, that the putative quadruplexes described are bound by the same molecules that bind other known quadruplexes, provides further evidence that they form quadruplexes *in vitro*, in addition to that from direct measurement of the oligonucleotides. It also provides initial insight into how they may be targeted *in vivo*.

2.7 *Cha* – an *in vivo* study of quadruplex activity

2.7.1 Introduction

Although the work described above, as well as previous work by other groups[59,125] has demonstrated that quadruplexes could be formed and bound, by small molecules or a protein, *in vitro*, the challenge proving *in vivo* function is considerably more challenging.[126] In order for these sequences to be candidate

targets for gene regulation, however, it is important to show that a physiological change can be produced using a quadruplex-binding ligand *in vivo*.

In order to measure the effect of a ligand, one approach is to establish a suitable reporter construct with a quadruplex in an appropriate position in the promoter region. Such a reporter can then be used to produce a specific and measurable change upon drug binding. An illustrative example of this process is shown below, taking as an example reporter Green Fluorescent Protein (GFP), which may be readily detected using fluorescence.

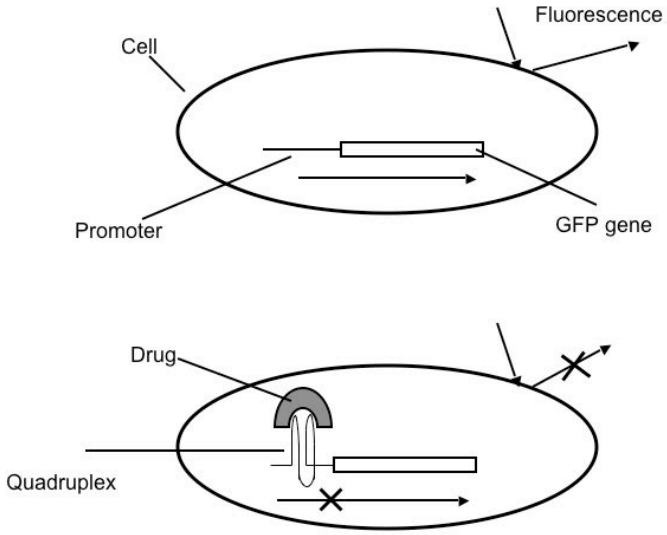


Figure 2.7.1. Example of *in vivo* assay for quadruplex binding and gene inactivation. Top: normal cell contains promoter of interest fused to GFP. Transcription is active and GFP is produced, giving an observable fluorescence. Bottom: drug addition induces quadruplex formation in the promoter, preventing GFP expression and eliminating fluorescence.

One organism used for a variety of genetic experiments is the fruit fly *Drosophila melanogaster*. It is favoured for such experiments because of its relatively short lifespan and the wide variety of mutants available.[158] One of the key genes for *Drosophila* neural development is the Choline acetyltransferase (Cha) gene.[159] This encodes a product with choline O-acetyltransferase activity involved in acetylcholine biosynthesis.[160] Acetylcholine is used by sensory neurons to stimulate interneurons and by interneurons to stimulate motor neurons. It is vital for the motor circuit, and without it larvae are paralysed.[158]

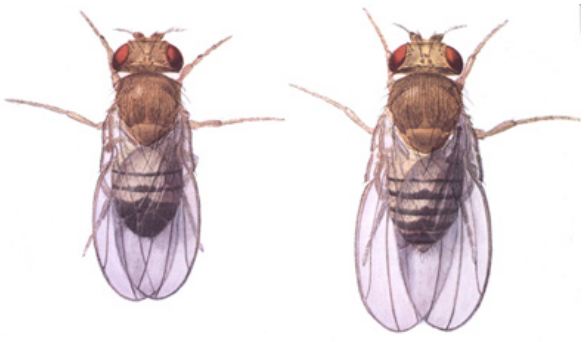


Figure 2.7.2. Male and female examples of *Drosophila melanogaster*, the common fruit fly.

Because of the importance of this gene, its promoter region is a subject of considerable interest, and a wide variety of mutant forms with varied promoters coupled to useful reporter genes are available, including those for GFP and lacZ. These allow expression levels for the *Cha* promoter to be coupled to observable fluorescence or colour changes.[158]

Analysis of the sequence of the promoter region of Cha reveals a putative quadruplex-forming site 182 bp before the transcription start site. It has the sequence shown below. Because this sequence occurs in a region known to be involved in gene regulation, and because of the ready availability of mutants suitable for our studies, it was considered as a useful model for *in vivo* studies.

Cha AATGGGCCTGGGAAAACGGGAAGCGGGCAA

In collaboration with Dr Marta Zlatic, a mutant strain of *Drosophila* has been used to test the activity of a number of known quadruplex binding drugs. This strain has the *Cha* promoter fused to the GAL4 gene, which then activates the UAS promoter fused to the LacZ gene. The aim of this project is to apply a variety of quadruplex-binding ligands to these flies as embryos, and to observe the effects of binding *via* the physiological effect of paralysis, and to use the LacZ reporter construct to confirm that the paralysis is due to underexpression of the Cha gene.

Three methods of drug application have been considered. The first of these involves mixing a solution of the test ligand with the food given to larvae, so that they ingest some amount of it. This has the advantage that it is very simple to

perform experimentally, but is not well controlled in terms of the amount of uptake by the larvae. Additionally, in order for any ligand to be active, it has not to be damaged by ingestion processes, and to reach the DNA target.

A second approach involves dissecting an embryo, and immersing it in a solution of the ligand being tested. This avoids the problem of ingestion, and the amount of ligand to which the embryo is exposed can be quantified accurately. However, it can be challenging to prevent the embryo from dying during the experiment, and it is still necessary for the ligand to enter the relevant cells.

A third approach involves direct injection of the ligand solution into the embryo at a very early stage in its development, such that distinct cells have not yet formed. This eliminates the problem of ensuring that the ligand enters cells, but does not automatically ensure that it will enter the nuclei. In addition, it is experimentally challenging to inject such solutions without killing the embryos.

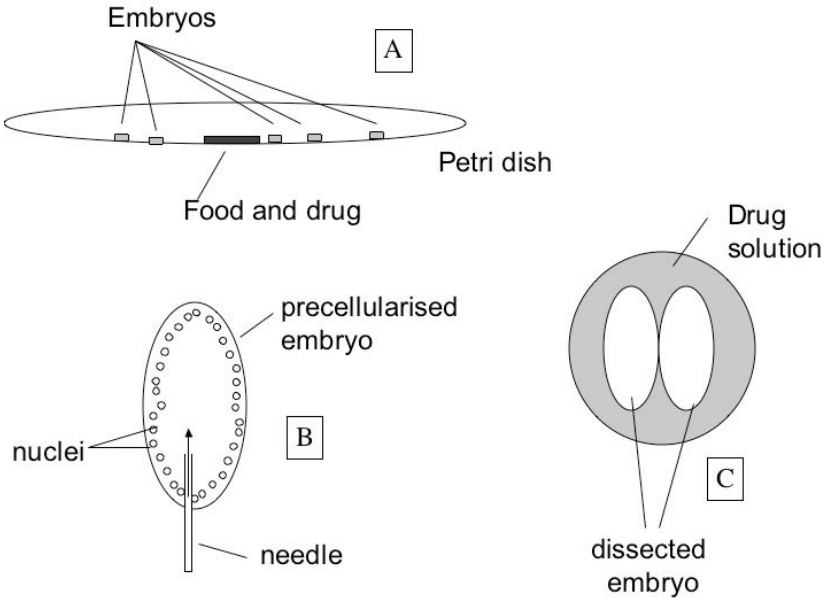


Figure 2.7.3. Three methods of administering drugs to fruitfly embryos. A) Drug solutions may be mixed with the food for the embryos so they ingest the drug. B) drug may be directly injected into precellurised embryos. C) embryos may be dissected and immersed in drug solution.

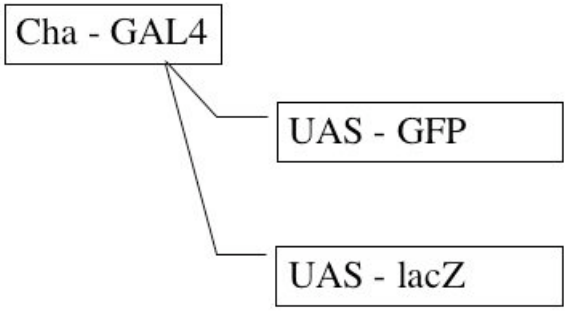


Figure 2.7.4. Promoter cascade for detection of Cha promoter activity. The Cha promoter is fused to the GAL4 gene, the product of which binds and activates the UAS promoter, which is fused either to GFP for fluorescent detection or lacZ for spectrophotometric detection.

In all of these cases, the bioavailability of the drug is a significant issue, and in the case of the ingested drugs, oral availability is also important. To address these issues, a selection of known quadruplex binders was modelled (using an algorithm online at www.molinspitation.com) to understand their bioavailability.

2.7.2 Biophysical studies on the Cha quadruplex

The cha quadruplex was studied by UV melting and CD spectroscopy. The CD spectrum shows a secondary structure, with a broad peak between 260 and 295 nm. This is not a classic signal for a quadruplex, but could perhaps arise from the superposition of a parallel and an antiparallel sequence. The UV melting results were much clearer, and show a clear hyperchromic transition,[129] with a melting temperature of 54 ± 1 °C.

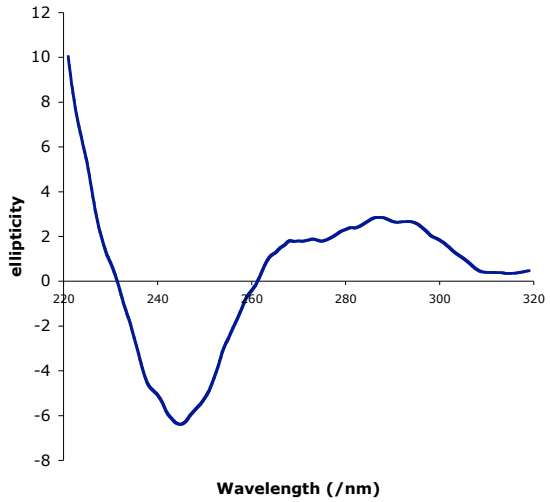


Figure 2.7.5. CD spectrum of the *cha* quadruplex. This shows peaks at 290 nm and 265nm, and could be due to superposition of two forms, one parallel, one antiparallel.

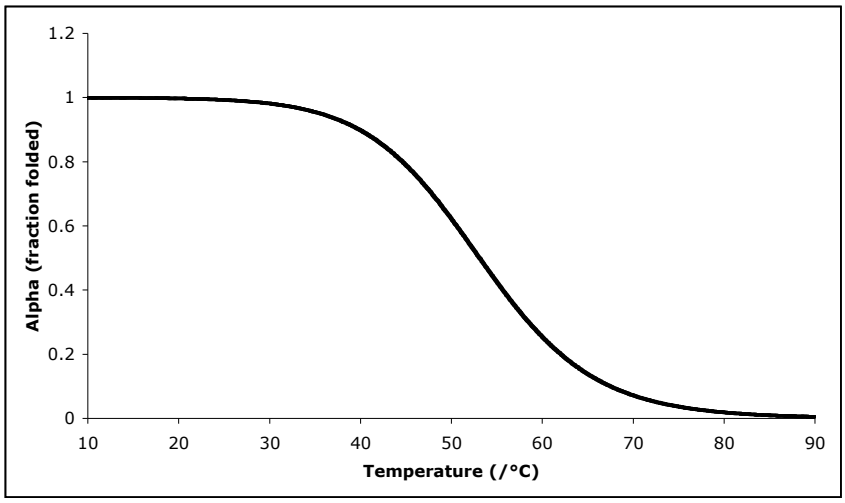


Figure 2.7.6. UV melting transition for the cha quadruplex. The UV data are shown as an alpha plot, showing the fraction folded against temperature, after correction for base line absorption changes. It shows a clear sigmoidal transition, with a midpoint at 54 °C.

2.7.3 Bioavailability

There are a number of factors affecting bioavailability, and accurate prediction of bioavailability is an important field for pharmaceutical development. One important factor is the hydrophobicity of the compounds, described as a logarithm, logP. This describes the partitioning constant between water and 1-octanol, where high values refer to lipophilic substances, and low values to hydrophilic ones. The values for novel compounds may be estimated by comparing the features of the molecule to those of experimentally measured compounds. The modelling software used, miLogP 1.2 by Molinspiration (www.molinspiration.com), was trained using many thousand drug-like molecules.

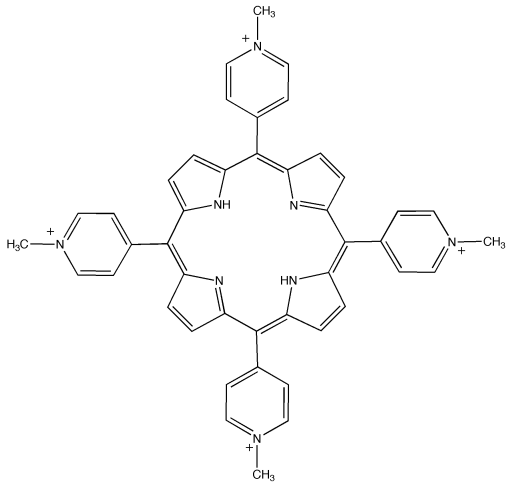
A more detailed approach developed at Pfizer and now used extensively is based on the Lipinski rules ('rules of five').[161] These are a set of criteria based on empirical observations. Drugs are more bioavailable the fewer of these rules they violate; almost 80% of bioactive compounds satisfying all four rules:

- The number of H-bond donors should be less than or equal to 5
- The number of H-bond acceptors should be less than or equal to 10
- The logP value should be less than or equal to 5
- The molecular weight should be less than or equal to 500

Although not part of the original Lipinski scheme, the number of rotatable bonds is frequently considered as well – lower values are favoured.

The compounds tested and the results found are shown below in figure 2.7.7. Also included for completeness is the protein Gq1, although small-molecule models are not appropriate for this protein, and so no values are given. The models used were also not able to calculate a sensible logP value for the charged porphyrin TMPyP4, and the value given is highly unreliable.

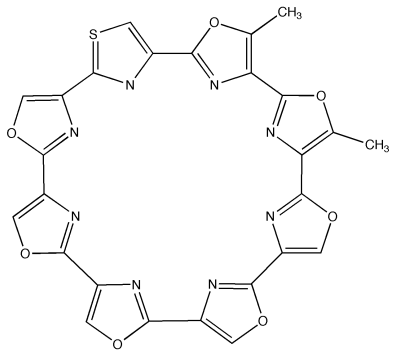
None of these drugs appears to be ideally suitable at this stage for development as drugs, with the exception of the cyanine dye. TMPyP4 is fairly good, with only one violation of the Lipinski rules. Clearly, further stages of drug development will be required to produce viable drugs.



TMPyP4

TMPyP4 was used for studies on c-myc,[125,126] and is readily available.

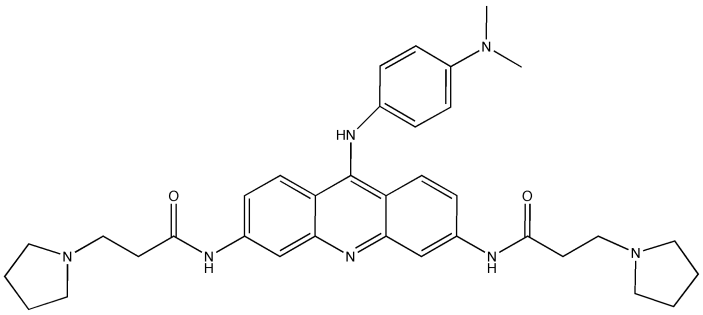
logP	-4.5 (?)	nON	8
PSA	72.9 Å^2	nOHNH	2
natoms	52	nviolations	1
MW	678.8	nrotb	4



Telomestatin

Telomestatin is the tightest known quadruplex binder[80]

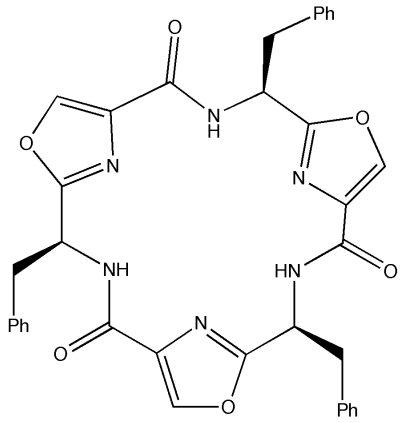
logP	5.1	nON	15
PSA	195.1 Å^2	nOHNH	0
natoms	42	nviolations	3
MW	580.5	nrotb	0



BRACO-19

Braco-19 is a highly specific binder used by the Neidle group.[79]

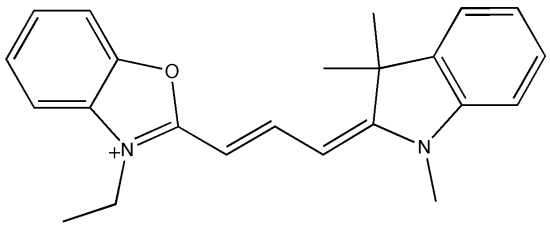
logP	5.8	nON	9
PSA	92.8 Å^2	nOHNH	3
natoms	44	nviolations	2
MW	593.8	nrotb	11



Triphenylalanine cyclic oxazole

This is one of a class of compounds being developed in our group by Katja Jantos as telomestatin mimics.

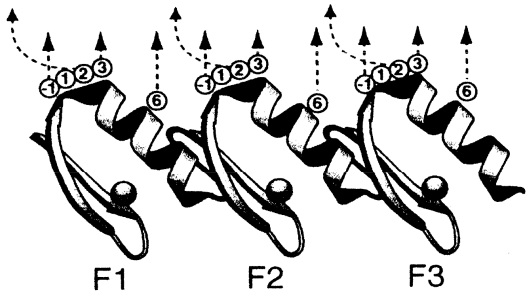
LogP	3.5	nON	12
PSA	165.4 Å^2	nOHNH	3
natoms	48	nviolations	2
MW	642.67	nrotb	6



Cy3 Ox vs. TMI

One of a class of cyanine dyes being developed in our group by Jen Hake.

LogP	2.8	nON	3
PSA	20.3 Å^2	nOHNH	0
natoms	26	nviolations	0
MW	345.5	nrotb	3



Gq1 – one of a family of three-zinc finger proteins developed in our group[92]

Figure 2.7.7 Structures and Lipinski modelling for various drug candidates. LogP measures the hydrophobicity, and should be below 5 for the drug to be available, but around 5 for it to be membrane-passable. PSA is the polar surface area, and alternative measure of hydrophobicity. Bioavailable compounds generally have PSA <140 $Å^2$. natoms is the number of atoms. MW is the molecular weight. NON is the number of hydrogen bond acceptors, and nOHNH the number of hydrogen bond acceptors. nviolations is the number of violations of Lipinski's four rules. Nrotb is the number of rotatable bonds. Calculations were performed using online software at www.molinspiration.com.

2.7.4 In vivo *tests*

The selection of ligands listed previously (except Gq1) was mixed at high concentration (millimolar) with larvae food, and the larvae were left to feed on it for 12 hours. All the embryos survived, but none exhibited any signs of paralysis or atypical GFP expression. Fluorescence examination of the larvae revealed that the drugs were present in the intestines, but not observably so elsewhere in the larvae. They had clearly not been broken down metabolically, as their fluorescence continued to be visible. However, because they were localised in the intestines they could not act on the cha promoter.

In an effort to avoid this problem, direct injection into embryos at the syncitial (precellular) stage, when there are no membranes around them has been attempted. So far, however, it has not been possible to keep the embryos alive after injections, using either drug injections or buffer injections as controls. This is a technique that has been reported in the literature, but in our hands has not yet been achieved.

A successful demonstration of proof of concept in such an advanced system would be a very powerful demonstration of the potential for the use of drugs to regulate gene expression *in vivo*. Such advanced experiments are being considered as future work in the medium term. We are also investigating other simpler organisms for study, and in particular study in single-celled organisms (eg bacteria) would make the studies simpler, if appropriate mutants could be prepared or otherwise obtained.

2.8 Conclusions

Quadruplex-forming sequences have the potential to be a powerful motif for both understanding and controlling gene expression and other biological processes, and for that reason are an interesting area of study. The first step in identifying physiological roles for quadruplexes is to develop a clear definition of what primary sequences have the capacity *in vitro* to fold into a quadruplex form. Clearly, a number of factors will have a bearing on this, and this chapter began by addressing the most significant of these in order to produce a folding rule, capable of predicting in general which sequences will form quadruplexes. This rule considers the presence and location of guanine bases to form the tetrad core of the quadruplex, and the length of the loops that join it, and makes a binary prediction regarding quadruplex formation. It is, however, only a first generation folding rule, and there are a number of factors it neglects, most importantly, it does not take account of the sequence of the loops, and does not predict any thermodynamic parameters. Considerable further work will be required in order to have a thermodynamic rule along the lines of the duplex model by Breslauer *et al.*[162], but it is nevertheless a powerful tool for examining sequences for the potential to form a quadruplex.

The simple nature of the folding rule makes it relatively straightforward to turn it into an algorithm, capable of reporting on the presence or otherwise of

potential quadruplexes in any piece of DNA (or RNA). Hence it is possible to identify individual regions of interest and examine them for potentially interesting quadruplex-forming sequences, and in this thesis a number of examples are given. This tool is used extensively in the studies described in the next chapter.

After identification of the sequences by bioinformatic methods, it is insightful to use biophysical techniques, such as UV spectroscopy and circular dichroism, in order to confirm their formation *in vitro*, and to develop an understanding of their thermodynamic properties. Eventually, it may be possible to predict many of these, but this is not yet possible. Detailed mutational studies and other reductionist approaches will play a significant role in enabling this to occur. X-ray and high-resolution NMR studies are of course the ultimate biophysical source of data on quadruplex formation, and this is being pursued in our lab in collaboration with Prof. S Neidle.

The last part of this chapter is concerned with an attempt to demonstrate this level of control *in vivo*, using a *Drosophila melanogaster* system. The fact that it has not been successful highlights the difficulties common to almost all *in vivo* systems, wherein the level of complexity is much higher than the reductionist *in vitro* approach.

CHAPTER 3

Quadruplexes in the genome

3.1 Introduction and rationale

Manual searching of interesting genes for putative quadruplex sequences (PQS) has the potential to reveal some interesting drug targets, but is a time-consuming process. Genome-wide searching allows much wider searches for PQS, as well as allowing statistical analysis of the presence of PQS in the human genome and in other genomes. It also facilitates higher-level searching of 'interesting' PQS, by allowing the combination of PQS position data with data for Single Nucleotide Polymorphisms (SNPs), conserved genes and many other pieces of genomic information.

In this chapter is described some fundamental studies on PQS in the genome, addressing for the first time the question of how many there may potentially be in the human genome. This frequency is then compared to that predicted if DNA was purely random (a Bernoulli sequence), and the reasons for the large excess found are discussed. The random frequency figures are derived both by Monte Carlo simulation and by explicit mathematical analysis.

DNA is, of course, non-random, and it is demonstrated that it may be better approximated as a Markov chain, with the frequency of each base being predicted by that previous to it. The extent to which genomic DNA differs from single-base randomness for individual chromosomes is discussed. This is still a simplification, as each chromosome has widely varying regions with different statistics.

In order to generate appropriately random control DNA, a windowing method is employed to make shuffled DNA with the bulk and local characteristics of genomic DNA, but no other conserved information. This is used to provide an improved estimate for the number of PQS that would be predicted in the genome.

As discussed previously, loop lengths play a significant role in quadruplex stability. If the PQS found in the genome are merely patterns with no

structural relevance, the length distribution for the three loops should be independent, and this distribution may be approximated mathematically as well as using Monte Carlo methods. On the other hand, if there are correlations between the loop lengths, as actually found, this is evidence for at least some of the sequences having a function in the quadruplex form.

In order to investigate which sequences are important, the extent to which they are conserved between different species is examined, and a number of well-conserved quadruplexes are identified. These are good candidates for having functionality of some form.

Another approach is to look for mutations that are known to be correlated with disease states involving over- or under- expression of genes. Where the mutations affect the stability of potential quadruplexes, these quadruplexes would also be good candidates for detailed investigation. An example affecting osteoporosis is discussed.

3.2 Search criteria and code

In order to search large amounts of DNA for putative quadruplex sequences (PQS), it is important to have an absolutely rigorous definition of what a quadruplex is. Throughout this work, the folding rule derived in the previous chapter is used. This rule may be expressed verbally as before:

‘Any sequence of the form d(GGG…GGG…GGG…GGG), where … represents a gap of between 1 and 7 bases (possibly including G), and GGG represents a series of at least 3 G residues, will form a quadruplex structure under appropriate conditions’

This may be expressed more rigorously as a regular expression of the form:

$$G_{\{3,6\}}N_{\{1,7\}}G_{\{3,6\}}N_{\{1,7\}}G_{\{3,6\}}N_{\{1,7\}}G_{\{3,6\}},$$

where N represents any base. The notation {i,j} refers to any integer between i and j inclusive. The {1,7} restriction for the loop lengths arises direct from the folding rule; the {3,6} restriction for the G-stretches arises from the need for at least 3 G residues being required to form a quadruplex, and the fact that with more than 6, the sequence could be expressed as (GGG) G (GGG), and so is already included as two repeats and a loop.

This definition is still at risk of miscounting sequences, however, as a sequence such as GGGTGGGTGGGTGGGTGGG could be aligned as:

(GGG) TGGGT (GGG) T (GGG) T (GGG)
or (GGG) T (GGG) T (GGG) T (GGG) TGGG

or many other ways. Any search program therefore has to have a shortest-possible loop assumption, in order to count each such quadruplex uniquely.

Another question to be tackled by such a search program is how to count degenerate quadruplexes, such as the example above. With five sets of GGG, this sequence could fold into five different quadruplexes by using four out of the five sets in each case (neglecting the shortest-loop assumption). So, should it count as five quadruplexes? Alternatively, should it be seen as two overlapping sequences, one with the left-hand four, and one the right-hand four? Or, should it be seen as one PQS only, as no more than one quadruplex could be formed at once using this DNA?

This situation becomes even more complex with longer strands – for example, if there are eight GGG repeats, that could be counted as ${}_8C_4 = 70$ (using the

combination approach), 5 (overlapping), 2 (maximum formable at once) or 1 (single continuous run).

In the randomised DNA sequences used as statistical controls in this work, the distinction is not very important, as long quadruplexes are rare events, but in genomic DNA these questions are important, especially with highly repetitive sequences such as telomeres. Each has merits in different applications, and it is important that a search program can count in each different way.

3.2.1 Quadparser – a rapid and flexible method for finding PQS

With assistance from Simon Rodgers of thaze, a programme was written to search genomic DNA very rapidly for PQSs. The source code is included in Appendix B. It can take in any DNA sequence in FASTA format.*

It is flexible in terms of the search criteria, which allow for easy variation of the allowed loop lengths, number of stacked guanines and other parameters, and also in terms of the output styles available, which can give simple counts, lists of chromosomes, gene identifiers and other options. It is written in C in order to be as rapid as possible, and can process the entire human genome (order 3×10^9 bases) in around 15 minutes on a 1.25 GHz G4 processor.

The input style is of the following form:

```
./quadparser <filename> <bases> <# bases in repeat> <repeats in sequence> <min gap size> <max gap size> <output file>
```

* FASTA format is a standard format for DNA and protein sequences. It involves a header, starting on a new line with a '>' sign. The next new line is the beginning of the sequence, with standard IUB/IUPAC codes for bases and nucleic acids. The end of the sequence is marked by a new line with a '>' sign.

Eg ./quadparser bases.fa G 3 4 1 7 output.txt will search through the FASTA file bases.fa looking for G-rich patterns ('G-patterns') that fit the folding rule (Four sets of three Gs, with loops of one to seven bases).

It can also search for equivalent patterns rich in the other bases, referred to herein as C-patterns, A-patterns and T-patterns (C-rich, A-rich and T-rich respectively). Although these have no physical meaning[*], the C-patterns signify a potential quadruplex on the complementary strand to that being studied, and is used because genomic DNA is normally only listed as the coding/sense strand. The A- and T-patterns are used only as statistical controls, because they do not relate to any known structure, and hence should show no biases from pure statistics.

In essence, the programme digests the FASTA header, and then goes through looking for sequences of the base being searched with the correct number of repeats (eg GGG). It then looks to see if the next set of repeats is within the allowed loop gaps, and if so repeats this until it has found the right number of repeats (normally four).

3.3 Genomic frequencies and distribution

3.3.1 Human genome analysis

A complete copy of the sequenced human genome map was taken from ENSEMBL, using NCBI build 34. Each chromosome was then individually *quadparsed*. A summary of the results is in table 3.3.1. Length and %GC contents are shown for the sequenced regions. There are a total of 3.1 billion bp in total, of which 91% has been sequenced.

Three counts are used for reasons discussed above. They are defined as follows, with explanations of how they would treat the octamer $(GGGN)_8$.

[*] In principle the C-patterns could form an *i-motif* structure, but this is not directly considered here.

Overlapping: Maximum number of quadruplexes that could in principle form from consecutive GGG runs. The octameric example would give 5.

Discrete: Maximum number of quadruplexes that could in principle be formed at any given time, with no overlaps. The octameric example would give 2.

Separate: Number of continuous sequences that could form any number of quadruplexes. The octameric example (like all examples) would give 1.

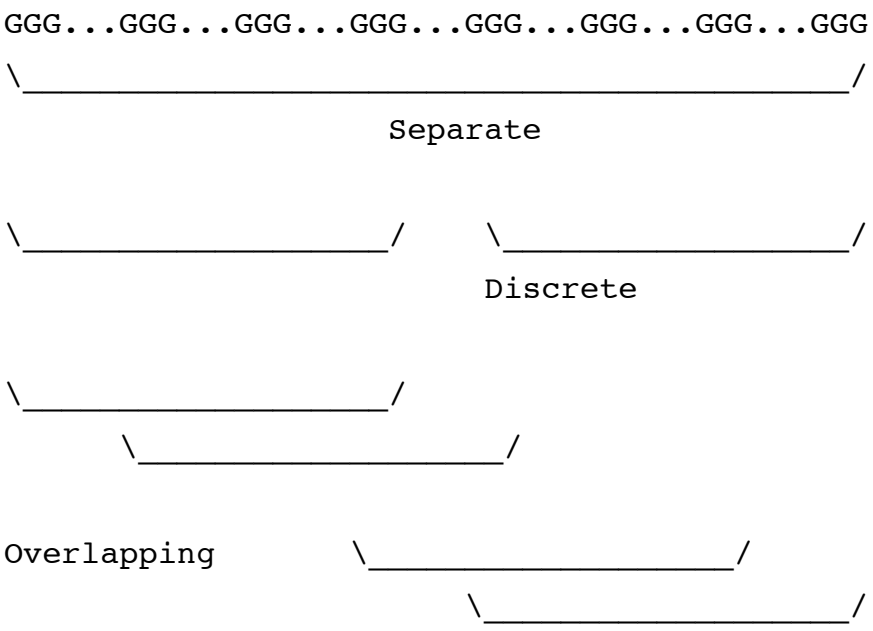


Figure 3.3.1. Three different ways of counting PQS. ‘Separate’ counts an entire string as one instance. ‘Distinct’ counts the maximum possible number of simultaneous quadruplexes that could form. ‘Overlapping’ counts the number of possible linear combinations that would lead to quadruplexes. ‘…’ represents a loop. The combinatorial options (counting any set of GGG that could form a quadruplex, even if not in order) have not been counted.

Chromosome	Length (Mbp)	%GC	Number of PQS (G or C-rich) found		
			Overlapping	Distinct	Separate
1	222	41.7	50.3×10^3	32.2×10^3	30.7×10^3
2	238	40.2	42.8×10^3	27.2×10^3	25.6×10^3
3	194	39.7	29.2×10^3	19.1×10^3	18.2×10^3
4	187	38.2	24.2×10^3	15.5×10^3	14.6×10^3
5	178	39.5	27.3×10^3	17.9×10^3	17.1×10^3
6	167	39.6	25.9×10^3	16.9×10^3	16.1×10^3
7	155	40.7	31.2×10^3	19.4×10^3	18.3×10^3
8	142	40.2	24.7×10^3	15.9×10^3	15.1×10^3
9	116	41.3	27.0×10^3	16.9×10^3	15.9×10^3
10	131	41.6	28.7×10^3	17.8×10^3	16.6×10^3
11	131	41.6	31.8×10^3	20.3×10^3	19.3×10^3
12	130	40.8	25.9×10^3	16.2×10^3	15.2×10^3
13	96	38.5	12.6×10^3	8.0×10^3	7.5×10^3
14	87	40.9	18.0×10^3	11.5×10^3	10.9×10^3
15	81	42.2	18.1×10^3	11.9×10^3	11.4×10^3
16	80	44.8	28.5×10^3	17.1×10^3	15.9×10^3
17	78	45.5	30.7×10^3	19.1×10^3	18.2×10^3
18	75	39.8	12.1×10^3	7.6×10^3	7.1×10^3
19	56	48.4	35.4×10^3	21.3×10^3	20.0×10^3
20	59	44.1	18.8×10^3	11.7×10^3	11.0×10^3
21	34	40.9	8.3×10^3	4.9×10^3	4.4×10^3
22	34	47.9	16.4×10^3	10.2×10^3	9.7×10^3
X	149	39.4	24.8×10^3	16.2×10^3	15.4×10^3
Y	22	39.0	2.3×10^3	1.6×10^3	1.5×10^3
TOTAL AND AVERAGES	2841	40.9	595×10^3	376×10^3	356×10^3

Table 3.3.1. Number of potential quadruplexes in each chromosome of the human genome by three different counts, along with length and %GC data. The three methods are described in the text, and would give counts for the sequence $(GGGN)_8$ of 5, 2, and 1 respectively.

A total of 595,000 quadruplexes is predicted, with a maximum of 376,000 at any given time. The decrease in the 'separate' method of counting from the 'distinct' method largely relates to a few highly repetitive sequences, such as telomeres and other such units. For comparison, the total number of A- and T-patterns is around an order of magnitude higher, due to the considerable AT-richness of the human genome.

This is clearly a very high number, and poses a considerable challenge for the validation of individual quadruplex targets. Fortunately, at any given time, it might be expected that only a small proportion of these are likely to be available for small molecule binding, the rest being tied as duplex sequences around histones in the nucleosome.[5]

In order to provide widespread access to the positional information of these sequences, the coordinates of every PQS in the human genome have been prepared in DAS format for visualisation *via* the ENSEMBL website, and are currently available for ENSEMBL displays in ContigView or CytoView forms from http://www.julianhuppert.org.uk/all_DAS.txt and are found on the enclosed CD.

In order to begin to address the question of whether these sequences could have biological significance, it is necessary to understand what the expected number would be in randomised DNA. This is the topic of the next sections.

3.3.2 Random Bernoulli DNA

The simplest way to study randomised DNA is to treat the DNA sequence as a series of independently-selected bases, with a certain probability of choosing each base. This has the advantage that as well as being easy to generate simulates for Monte Carlo analysis, it is mathematically tractable.

An approximation to this probability comes from Staden, who showed[163] that the probability of a string such as the one for PQS with variable gaps has a probability equal to the product of the probability for any one matching pattern,

multiplied by a factor of n for each gap, where n is the number of possible values for the gap length. This then gives an estimate for the quadruplex density of

$$\begin{aligned}\rho(PQS) &= n^{\text{no. loops}} \cdot p(GGG.GGG.GGG.GGG) \\ &= n^3 \cdot p^{12} \\ &= 7^3 p^{12} \quad = 343\, p^{12}\end{aligned}$$

where *p* is the probability of any given base being a guanine. However, this is not a complete treatment, and as will be seen does not well describe the actual probability, although it does correctly predict the leading term.

Sewell and Durbin[164] developed an algorithmic method for estimating the probability of finding bounded regular expressions in random data, intended specifically for identifying protein motifs. This could in principle be applied to the current problem, but due to the repeated structure of the quadruplex form, would be relatively inaccurate and slow, especially since an exact calculation is possible in this case.

The full mathematical treatment is given in full in appendix A, and derives an expression for the density of quadruplexes in a single strand. The counting method relevant to this treatment is the 'distinct' method, as the method does not allow for overlapping sequences (although it could in principle be extended to cover these, by treating this as a Poisson process).

In essence, there must be a 12^{th} order term in *p*, the probability of any given base being a guanine, due to the requirement of having 12 guanine bases in relatively defined locations. There is then a cubed term in *p* to describe the frequency of finding three appropriate loops. This term must be larger than 1, as the tolerance of varied loop lengths increases the probability of finding the correct sequence.

The final result is a polynomial with terms in p^{12} through p^{27}. Since *p* is relatively small (around 0.2 for most genomic sequences), the terms in lower

orders of *p* are dominant, with the quadruplex density $\rho(PQS)$ expressible as the truncated sum:

$$\rho(PQS) = 343p^{12} - 882p^{13} + 756p^{14} - 1098p^{15} + ...$$

It may be seen that the simple approach of Staden[163] does give the correct first term, but because the sequence does not terminate rapidly, misses a lot of important information.

To confirm that this equation was correct, a number of Monte Carlo simulates were produced with values of *p* ranging from 0 to 0.4. These results are shown below, along with a least-squares fit to a polynomial in orders 12 through 15.

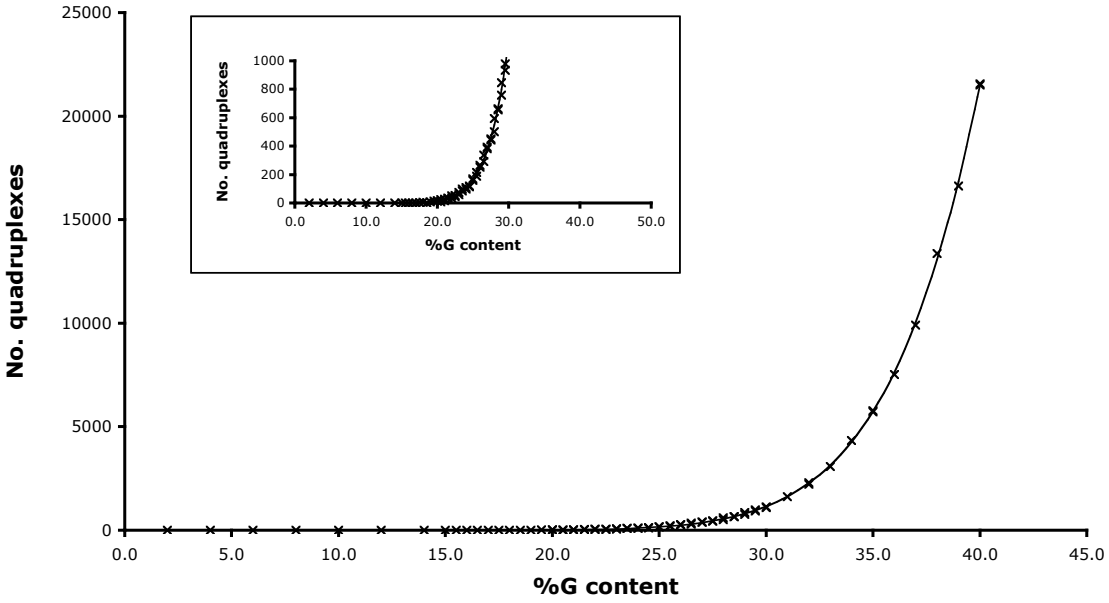


Figure 3.3.2 Monte Carlo simulations. 20 Mb chromosomes with variable proportions of guanine were randomly generated and *quadparsed* using the 'discrete' counting method. A polynomial with terms from p^{12} to p^{15} was fitted to the resulting counts by least-squares minimisation. Inset: expansion of the low-quadruplex-forming region to show fit quality.

The exact polynomial terms used to fit the equation have a considerable effect on the quality of the agreement between the coefficients from the Monte Carlo fit and the theoretical calculation. This is largely due to the fact that although the terms of the theoretical result decrease as the order increases, the rate of

decrease is relatively slow, and so the truncation is not very accurate. To demonstrate this, the following table shows the results from a least-squares fit for various orders, with the coefficients shown.

	Coefficients					
No. terms	p^{12}	p^{13}	p^{14}	p^{15}	p^{16}	p^{17}
2	234	-424	-	-	-	-
3	309	-839	567	-	-	-
4	309	-867	725	-215	-	-
Calc'd	343	-882	756	-1098	+2835	-3357

Table 3.3.2. Coefficients from fitting polynomials of 12th order and above to Monte Carlo data for 'discrete' quadruplex frequency in 20 Mbase simulated DNA with various frequencies of guanine. The theoretical coefficients are shown as derived in Appendix A.

This shows that the coefficients are close to those calculated, as long as sufficiently more terms later in the polynomial are taken into account. This problem is particularly acute because of the alternating nature of the signs in the polynomial expansion.

Interestingly, the Monte Carlo results also fit extremely well to a simple power rule, with $\rho(PQS) \propto p^{10}$ (see fig 3.3.3), but not at all well to other power rules, such as Staden's p^{12}. Despite the apparently good fit, this tenth power has no theoretical rationale, and can be seen to deviate for small p (inset A below), or by plotting $\rho(PQS)/p^9$ against p (inset B below) – if the power law were the correct form of the relationship, this plot should be linear through the origin.

Using this relationship and the observed frequencies of guanine for each chromosome, it is possible to calculate the predicted number of quadruplexes in the human genome assuming there were no selective pressures for or against quadruplexes. The other assumptions are that the DNA is a random Bernoulli stream, and that each chromosome is homogenous. The results are given in table 3.3.3

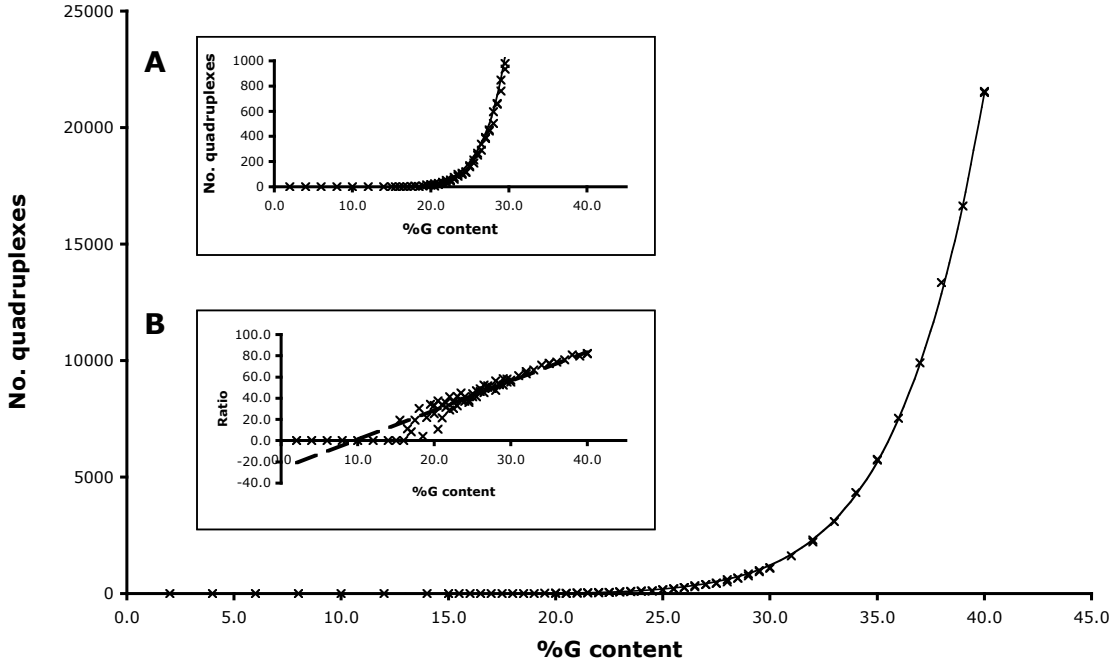


Figure 3.3.3 Result of fitting Monte Carlo simulations to a simple 10^{th} order power law. The fit appears to be extremely good, but deviates slightly for low G-densities (Inset A). Inset B: Plot of (No. quadruplexes)/(%G^9) against %G. If the power law were correct, this would be a straight line through the origin. A line of best fit is shown as a guide – it does not pass through the origin. The values below %G = 15 are subject to quantisation, as frequently no quadruplexes at all were found in the 20 Mbases simulated.

	G-pattern	C-pattern	A-pattern	T-pattern
Modelled	4,177	4,150	150,634	152,977
Observed	188,836	187,610	1,624,670	1,734,285

Table 3.3.3. Modelled and observed numbers of quadruplexes in the human genome, assuming that DNA is a random stream. The 'discrete' counting method is used.

It is clear that there is a very much larger number of actual PQS than that predicted by the model. There are two possible explanations for this. The first is that there is a selective pressure in favour of quadruplex-forming sequences. The second is that the assumptions made in the model are not appropriate. The former cannot be concluded from this data, given that many more AT-quadruplexes are also found than predicted. Since these have no known physical meaning, it would be hard to explain a selective pressure in their favour. Hence we must re-examine the assumptions made.

The two key assumptions in the model were that DNA may be treated as a Bernoulli stream, with each base independent of those adjacent to it, and that each chromosome is homogenous. In the next section, I investigate whether DNA can be well approximated by the Bernoulli method. The section after that develops a windowing approach to chromosomes, to deal with inhomogeneity along the chromosomes.

3.3.3 Diad frequencies

The simplest model for DNA is as a stream of independent bases. The next level of complexity is to consider the DNA as a Markov chain,[165] where each base is dependent solely on the base before. Thus the frequency of obtaining a G in position n may be different with different bases in position n-1.

Does DNA behave in a Markov sense? To investigate this, each chromosome of the human genome was subjected to an analysis in which each possible diad arrangement was counted, as well as a simple base count. As an example, the results from chromosome 1 are shown below, both as raw frequencies and as departures from the probabilities predicted using the single-base frequencies alone. Similar studies have previously been performed.[166]

Next base	**Previous base (chance, deviation from random)**								
	G		**C**		**A**		**T**		**(total)**
G	26	+4	5	-16	24	+3	25	+4	21
C	21	-	26	+5	17	-4	21	-	21
A	29	-	35	+6	33	+4	22	-7	29
T	24	-5	34	+5	26	-4	33	+4	29

Table 3.3.4. Diad analysis of human chromosome 1. Vertical lines show the percentage chances of each base following a given base, and then the deviation from the value expected if each base was independent.

It is immediately apparent that the diad base frequencies are significantly different to the frequencies obtained for individual bases. Analysing the whole genome gives a consistent picture, as shown in the table below, which shows the average and standard deviation of these figures for every chromosome.

	Previous base (chance, standard deviation)								**(total)**	
Next	**G**		**C**		**A**		**T**			
base	**%**	σ	**%**	σ	**%**	σ	**%**	σ		
G	<u>25.7</u>	1.7	(5.0)	1.0	<u>24.1</u>	1.9	<u>25.0</u>	1.7	20.7	1.3
C	21.1	1.1	<u>25.8</u>	1.7	(17.3)	0.9	20.4	1.3	20.8	1.3
A	28.7	1.4	<u>35.2</u>	1.5	<u>32.7</u>	1.3	(21.7)	1.7	29.2	1.3
T	(24.5)	1.4	<u>34.0</u>	1.1	(25.8)	1.5	<u>32.8</u>	1.3	29.3	1.3

Table 3.3.5. Diad analysis of every human chromosome. Vertical lines show the percentage chances of each base following a given base, and then the standard deviation across the chromosomes. Significantly 'enriched' probabilities are underlined. Significantly 'depleted' probabilities are enclosed in brackets.

It is clear that there are strong patterns in these data. Two main threads may be observed: Homodiads are significantly more common than expected, and CpG sequences are highly disfavoured. The former could perhaps be explained by suggesting that polymerases may be prone to repeating bases, but the latter observation is harder to explain, especially as it is so strongly asymmetric. Previous observers have ascribed this either to higher genetic mutation rates of CpG sequences, or the fact that CpG methylation is used as an important epigenetic code, and so it is selected against in regions where such labelling is not desired.[167]

The prevalence of homodiads, however, will mean that quadruplex-like sequences, having a core of homodiads, will be significantly more common than would otherwise be expected, and so it is necessary to take account of this property in developing control sequences.

3.3.4 Windowed DNA

Since chromosomal DNA is not homogenous, but consists of regions of highly variable base composition, it is not appropriate to model it with a single structure. The reasons for this variable composition are that different areas of the genome have different roles to play, either grossly or more locally. For example, introns and exons have different compositions because the latter is constrained by the amino acids it will form, whereas the former is not. Repeated regions such as telomeres and centromeres will have their own specific characteristics as well. Picture 3.3.4 shows ENSEMBL data for human chromosome 21, showing some of the variability in GC-richness, repeat frequency and gene presence along the chromosome.

As an example of the effect of this non-uniformity, consider a hypothetical minichromosome of 1 kb, which is 10% G-rich. If those 100 G's are located reasonably uniformly throughout the chromosome, the chance of finding a quadruplex would be extremely low. On the other hand, if they were all clustered in the first 150 bases (as, say, in a telomere), the chance of finding some becomes extremely low.

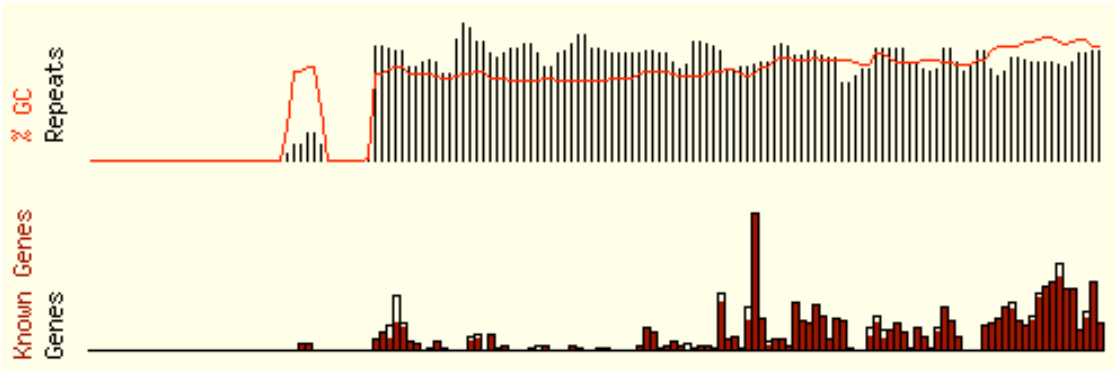
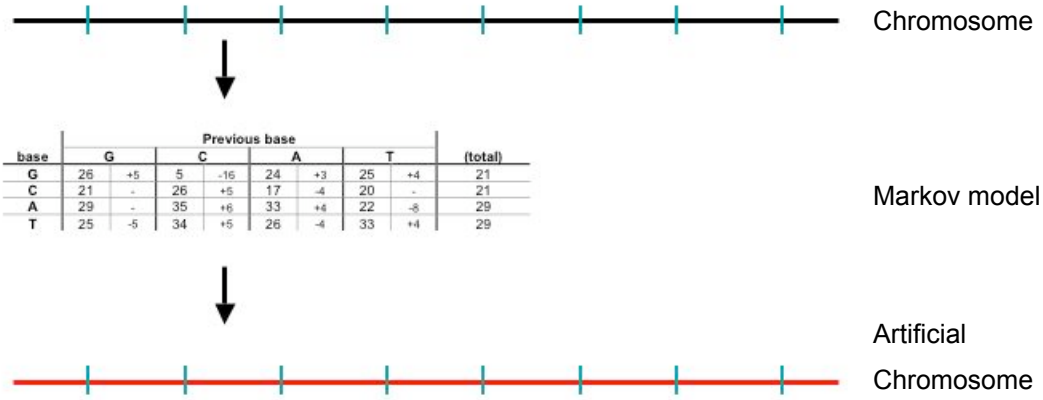


Figure 3.3.4. ENSEMBL data for chromosome 21 (NCBI build 34). The aligned plots show % GC content (red line), repeat frequency (black bars), and gene locations (black boxes, red if known other than solely computationally). Regions with apparently no GC content (left) have not yet been sequenced.

An alternative treatment, which deals with this problem reasonably well, is to break the chromosomes into small windows, and to make only the assumption that the DNA is homogenous along this small stretch. This can be used to simulate DNA with either a simple probability model or the diad model discussed above. The process may explained as in figure 3.3.5:



	Previous base								
base	G		C		A		T		(total)
G	26	+5	5	-16	24	+3	25	+4	21
C	21	-	26	+5	17	-4	20	-	21
A	29	-	35	+6	33	+4	22	-8	29
T	25	-5	34	+5	26	-4	33	+4	29

Figure 3.3.5. Generation of diad windowed simulates. The real chromosome (black) is divided into windows, each n bases long (green dividers). Each window is then analysed to produce a Markov model (table). This model is then used to generate n bases of artificial chromosome (red), and the process is repeated.

The length of the window is an important parameter in determining how effective this treatment is; if it is too long, then inhomogeneity in the real chromosome is lost – in our case, this will result in modelling fewer quadruplexes than expected. On the other hand, if it is too short, then the resulting sequences will diverge considerably from the real sequences, as they are more prone to statistical errors. In the extreme, of course, 1-base or 2-base windows will result in a perfect replica of the starting model – not a useful control sequence!

Using a window with an initial arbitrary size of 200 bases, I have generated 5 scrambled versions of each chromosome. These have the same diad statistics as the real chromosomes. Using *quadparser*, I have then counted the number of quadruplexes in each replicate, allowing me to develop a model for the number of quadruplexes that would be expected from this model, along with a standard deviation (below 1%). These results, along with the real values and those using the naïf model described previously, are shown in table 3.3.6. The total number of G- and C- patterns have been summed, as have the A- and T- patterns.

Method	No. G- and C- patterns	No. A- and T-patterns
Real chromosomes	376 x 10^3	3,359 x 10^3
Simple Bernoulli	8 x 10^3	304 x 10^3
Markov Windowed	269 x 10^3	2,019 x 10^3

Table 3.3.6. Number of G- and C-patterns, and A- and T-patterns, found using three different methods. The top row shows the actual number found in the human genome. The second row shows the number modelled for artificial chromosomes generated considering only the individual base frequencies. The third row shows the number in artificial chromosomes using the Markov method in 200-base windows of the real chromosomes. The 'discrete' method of counting quadruplexes is used.

It is clear that the diad windowing model gives results that are significantly closer to those found in reality for GC-quadruplexes than for the simple Bernoulli model. However, it does introduce the artificial parameter of the window size into the model, and there is no *a priori* argument that can account for what that value ought to be. For that reason, I generated a series of scrambled chromosomes using the Markov windowed method outlined above, using window sizes ranging from 50 to 4000 bp. The results are shown in table 3.3.7, along with the real results and outcome of the Bernoulli simulations.

These results show that there is a very strong dependence on the window size in terms of the number of each type of pattern identified. The number of patterns predicted increases almost hyperbolically as the window size decreases towards zero, as shown in figure 3.3.6. In order to select an appropriate window size, therefore, it is necessary to make some assumptions to calibrate the model. The assumption that was made is that AT-patterns would not necessarily be subject to the same pressures as the GC-patterns, because they do not have a propensity to form secondary structure. By making this assumption, it is possible to calibrate the window size to a value that gives the correct prediction for the AT-patterns. This occurs for a window size 75 bp, when the model correctly predicts 3.26 million AT-patterns (see table 3.3.7).

Method	GC-patterns	AT-patterns
Markov, size 50	687×10^3	4.01×10^6
Markov, size 75	514×10^3	3.26×10^6
Markov, size 100	420×10^3	2.81×10^6
Markov, size 150	320×10^3	2.29×10^6
Markov, size 200	269×10^3	2.02×10^6
Markov, size 400	185×10^3	1.56×10^6
Markov, size 1000	123×10^3	1.20×10^6
Markov, size 2000	93×10^3	1.02×10^6
Markov, size 4000	75×10^3	0.89×10^6
Bernoulli	8×10^3	0.30×10^6
Real human genome	376×10^3	3.26×10^6

Table 3.3.7 Total number of GC- and AT-patterns found in the real human genome and simulates using various methods. In the window methods, simulates were generated conserving diad base frequencies in windows of the size shown. Five independent repeats were performed, and the standard deviation was in all cases less than 1%. The 'Bernoulli' method treats DNA as a stream of independent bases, with base frequencies homogenous across each chromosome.

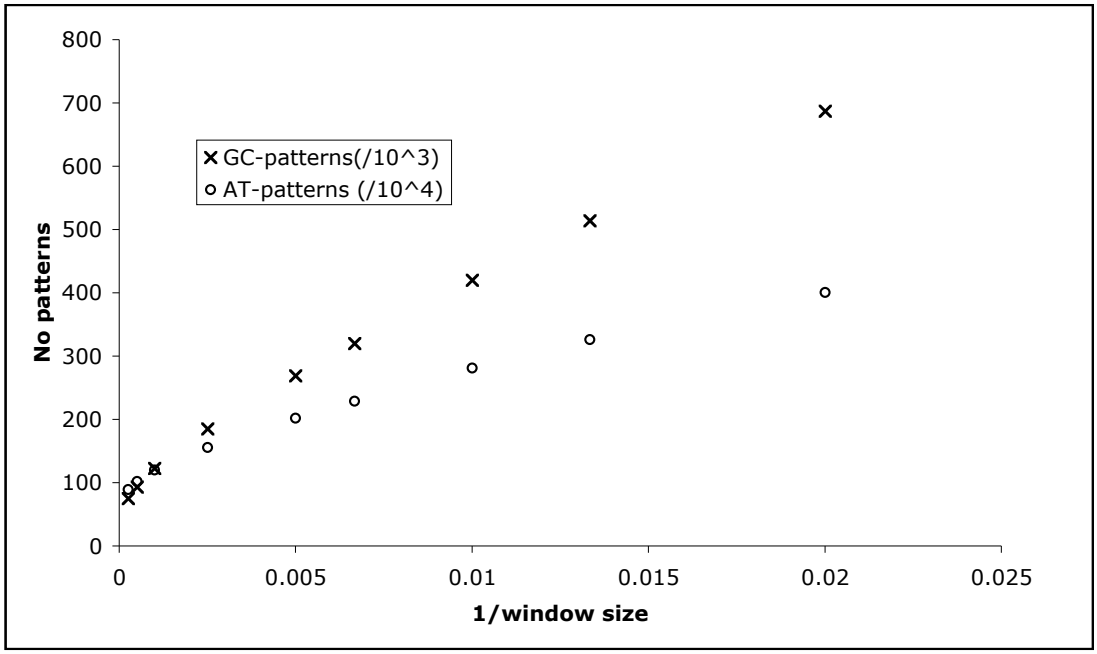


Figure 3.3.6 Plot of no. patterns found against the reciprocal of the window size. For low values of the window size (<1000) this appears to be linear. GC-patterns are shown in units of 1000, AT-patterns in units of 10,000.

Using this window size of 75 bases, the predicted number of GC-patterns is 514,000, as against an actually observed number of 376,000. Hence the model predicts 37% more GC-patterns than are actually present. Using smaller window sizes over-predicts both GC- and AT-patterns, windows of around 100 bp over-estimate the GC-patterns but under-estimate the AT-patterns, and above 150 bp, both species are over-estimated. The difference in the ratios is shown in a different format in table 3.3.8, as ratios of each type of pattern, normalised to n(T-patterns) = 1. Small variations in the detail of the folding rule, such as restricting loop lengths to be between 2 and 6 bases, give essentially similar results.

It is worth highlighting that for all of these results, there is a pseudo-Chargaff rule, with the number of G-patterns equalling the number of C-patterns, and the same for A and T. This arises from the underlying symmetry of the two strands in a chromosome, where the choice as to which strand is listed is purely arbitrary.

	Relative frequencies of X-patterns, normalised to T=1 within each column										
X	Actually observed	Bernoulli	Markov 50 bp	Markov 75 bp	Markov 100 bp	Markov 150 bp	Markov 200 bp	Markov 400 bp	Markov 1000 bp	Markov 2000 bp	Markov 4000 bp
G	0.12	0.03	0.17	0.16	0.15	0.14	0.13	0.12	0.10	0.09	0.08
C	0.12	0.03	0.17	0.16	0.15	0.14	0.13	0.12	0.10	0.09	0.08
A	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.99
T	1	1	1	1	1	1	1	1	1	1	1

Table 3.3.8: Frequencies of X-patterns for x = G,C,A,T under various conditions, normalised such that the frequency of T-patterns in each column is 1. The actually observed data is taken from the NCBI build 34 of the human genome. The Bernoulli model and Markov models are described elsewhere, and the number below the word 'Markov' refers to the window size used. It can be clearly seen throughout this data that the pseudo-Chargaff rule G=C and A=T holds. This also shows that the relative depletion of GC-patterns increases with increasing window size.

A 37 % decrease in the number of PQS is a significant observation, and suggests that these patterns have at least some physiological relevance. It also suggests that any evolutionary pressures that have occurred have served in general to disfavour PQS formation. Perhaps the quadruplex secondary structure disturbs processes such as DNA replication?

3.3.5 Location of PQS in the genome

So far, the genome has been treated as consisting of undifferentiated regions, and features such as coding regions have been ignored. However, we may expect that if quadruplex-forming sequences play a physiological role, their prevalence may vary depending on the physiological function of their location. Using the ENSEMBL database, it is possible to classify regions of the genome as related to genes. This has been used to investigate the number of putative quadruplexes in genes, and specifically within the exonic regions. These results show marked differences in terms of base composition and frequency of quadruplex-forming patterns. The exonic regions have been determined using the annotations within ENSEMBL. They have a base ratio that closely approximates equality (G:C:A:T 1.05:1.04:1.07:1), and diad repeat ratios that deviate less far from a Bernoulli model than the rest of the genome (GG:CC:AA:TT 0.98:1.07:1.05:1 for exons, against 0.78:0.79:1.00:1 for the whole genome). There are still some differences in these parameters between the four bases, and as a result they give rise to different predicted frequencies, based on a windowed Markov model of the exons, as previously described. These ratios are G:C:A:T 0.91:1.10:1.21:1. Examination of the actual exonic regions reveals a very different ratio, with the observed frequencies being in the ratio G:C:A:T 0.48:0.83:0.93:1.

Base	Frequency	Diad repeat Frequency	Observed pattern frequency	Predicted pattern frequency	Observed/predicted frequency ratio
G	0.25	0.27	0.48	0.91	0.53
C	0.25	0.29	0.83	1.10	0.75
A	0.26	0.29	0.93	1.21	0.77
T	0.24	0.27	1	1	1

Table 3.3.9: Frequencies of bases, diads and patterns for each base in the exonic regions of the human genome. Frequency lists the frequency of each bases in the relevant region. Diad repeat frequency refers to the chance that after a given base, the same base will be repeated. The observed and predicted pattern frequencies refer to the relative frequencies of patterns of the form $d(X_{3+}N_{1-7}X_{3+}N_{1-7}X_{3+}N_{1-7}X_{3+})$ for X= G,C,A,T, normalised to 1 for the frequency of T-patterns in each column, either in the actual human genome or in a simulate using a Markov model with a window size of 75 bp. The data shows that the G-patterns are dramatically underrepresented, and there is a weaker effect on C-patterns, and another on A-patterns.

One feature of these observations is that the pseudo-Chargaff rule that applies for the whole-genome, does not apply to the case of exons. This is unsurprising, since the two strands of DNA are distinct in exonic regions, with only one of them being transcribed to form RNA, which will have the same sequence (post-processing) as the other strand. However, does this show a bias with regard to quadruplex location as well? To investigate this, the differences between predicted ratio of the X-patterns, compared to the observed ratio, were considered. This is given in table 3.3.8, and gives a ratio of observed/predicted of 0.53, 0.76, 0.76, 1 for G, C, A, T.

The most marked effect here is the considerable suppression of the G-patterns compared to that predicted (0.53 observed/predicted). This effect is much more marked than that observed for C-patterns (0.75), which is of a similar order to the effects on A-patterns (0.77). Why is this difference observed? The simplest explanation is that since the G-patterns would lead to potential quadruplexes in the mRNA strand or the DNA duplex, whereas C-patterns could only lead to quadruplexes in the DNA, this is evidence of evolutionary pressure to reduce the number of quadruplexes allowed to form in mRNA. This might be because quadruplex formation in mRNA prevents translation from occurring, shortens half-lives, or in some other manner interferes with the normal role of mRNA.

To date, there has been relatively little work focused on RNA quadruplexes, although it has evoked some interest.[116,168] The observation that there are statistical differences in the pressures acting on mRNA-forming and non-mRNA-forming sequences suggest that where such quadruplex-forming sequences can exist, they may have some function, and should be investigated further; our laboratory is beginning such studies.

3.3.6 Loop length survey

3.3.6.1 Length distribution

A distinct approach to investigating whether these putative quadruplex sequences have experienced selection pressures is to investigate the distribution of loop lengths found in genomic DNA as opposed to simulated DNA. To investigate this question, counts have been made of the lengths of each loop for the quadruplexes found. Where there are several possible quadruplexes, each one has been counted (so a sequence of the form $GGGN_1GGGN_2GGGN_3GGGN_4GGG$ would be counted as loops of length {1,2,3} and {2,3,4}. Initially, a simple analysis was performed, with histograms being produced for each loop independently, describing the number of instances of each loop being of length 1,2, … 7.

This study has been performed both on the real human chromosomes and on simulates of chromosome 1, generated using the Markov model described previously, with a window size of 75 (10 replicates were performed, and the average result used. The differences between replicates were very small. The results are shown in figure 3.3.7. The plots on the left show the raw histogram data for the three loops. In every case, the frequency of loops with length 1 is very high, as predicted by the mathematical model in the appendix, and then gradually decreases with increasing loop length after length 2. However, on close inspection it can be seen that for the entire human genome, the first and third loops have the same frequency distribution, but the second loop does not behave exactly the same as the first and the third, deviating slightly from them, especially for loop lengths 3,4, and 6. This is not observed in the simulates.

In order to graph a more significant representation of this difference, the percentage excess (or decrease) of the frequency at any loop length of loop 2 over the average of loops 1 and 3 was calculated, and is shown in the plots on the right of figure 3.3.7. The formula used is shown below. This shows a clear excess of sequences with the central loop being lengths 3 or 4, with a

corresponding decrease in long lengths for the central loop. Again, no significant effect is observed in the simulates.

$$\%\ \text{excess, length x} = 100\frac{n_{2,x} - \frac{1}{2}\left(n_{1,x} + n_{3,x}\right)}{\frac{1}{2}\left(n_{1,x} + n_{3,x}\right)} = 100\left(\frac{2n_{2,x}}{n_{1,x} + n_{3,x}} - 1\right)$$

where $n_{a,b}$ refers to the number of times loop *a* had length *b*

Since in duplex DNA the 'loops' have no significance, it would be expected that they would all have the same statistical distribution. In contrast, in a quadruplex folded structure, loop 2 would be in a structurally different position from loops 1 and 3. The fact that we do see a significant (8%) discrepancy between the central and the side loops in the data from the actual human genome means that at least some of the PQS sequences observed may have been subject to a selective pressure based on their quadruplex role.

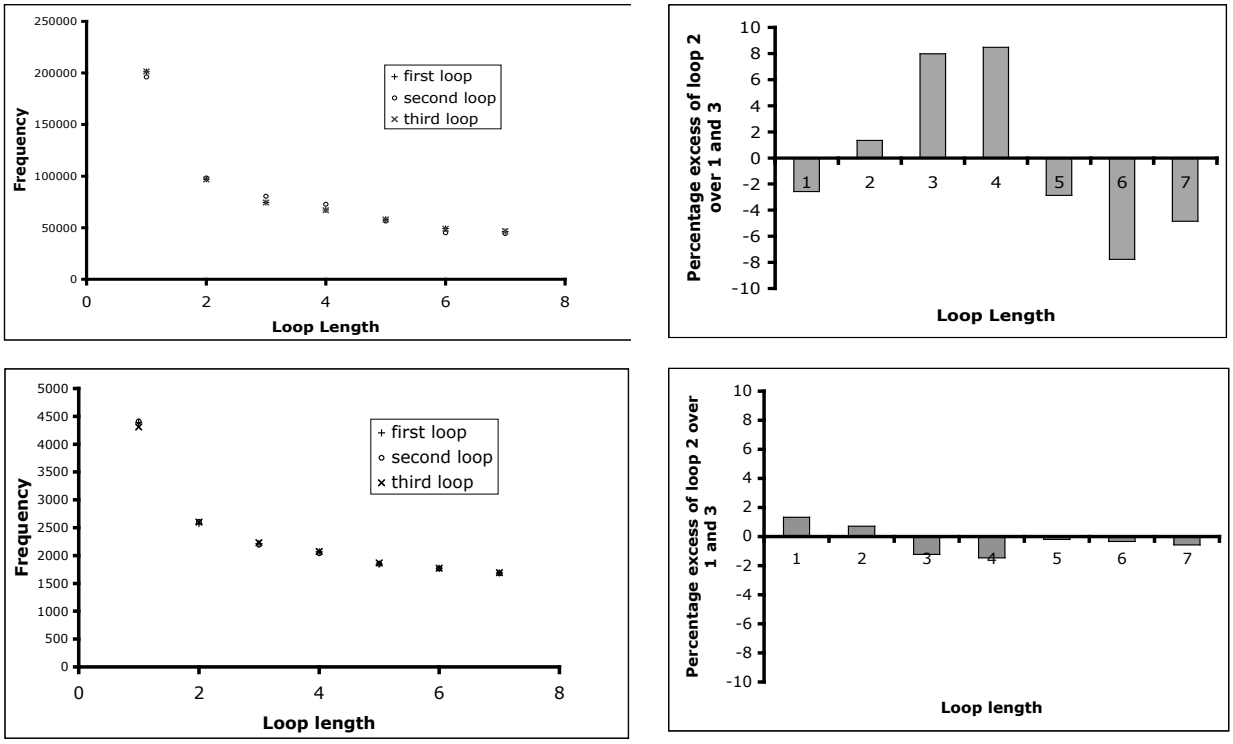


Figure 3.3.7. Left: Frequency distributions of loops of lengths 1-7 bases for the entire human genome (top) and diad windowed simulates of chromosome 1 (bottom). Right: percentage excesses of loop 2 counts over the averages of loops 1 and 3 for the entire human genome (top) and simulates of chromosome 1 (bottom). Chromosome 1 simulates are shown as the average of 10 independent repeats.

3.3.6.2 Length correlations

In order to further develop this approach, a slightly more detailed statistical approach was taken, using the ~3500 quadruplexes in the repeat-masked sequence of chromosome 21 as an initial sample. (This chromosome was selected because it is a particular focus of research at the Sanger Institute and their mapping projects). The data were repeat-masked to remove low-complexity regions such as telomeres, as they would be expected to strongly bias the results. A set of triplet values representing the lengths of each loop for a PQS was produced using *quadparser*, and these analysed using the program MINITAB. There was a very weak but statistically significant positive correlation between the lengths of each loop. I suspected that there may be subcategories of quadruplexes within this data, due to the various folding patterns adopted, and that the result of these separate sets of correlations gave rise to a small correlation overall. To investigate this, the data were split into two roughly equally sized categories, based on the anticipated folding patterns. It should be highlighted that the boundary between the two categories is not absolutely clear-cut, as the folding pattern has a large number of determinants aside from just loop length.

3.3.6.3 Antiparallel folds

As shown in chapter 2, when the loops of a sequence are relatively long, the most stable fold is generally an antiparallel form, with loops 1 and 3 on one side, and loop 2 on the other side of the quadruplex core. Such structures would have the potential for interactions between the bases of loops 1 and 3, as found experimentally in many sequences, such as the NMR solution structure of the human telomeric repeat.[27] Figure 3.3.8 shows this structure with the loops highlighted. Loops 1 and 3 are seen to interact *via* hydrogen-bonding and stacking interactions, but loop 2 is remote and cannot interact with the other loops.

If this interloop interaction is a stabilising interaction, and if selective pressure favours stable complexes, it might be expected that there is enrichment for sequences with loops 1 and 3 of similar lengths. In order to test this, the

subset of quadruplex loop data where all loops were 3 or more bases long (in order to select antiparallel folds) was examined using Pearson correlation analysis[169] on each pair of loops.

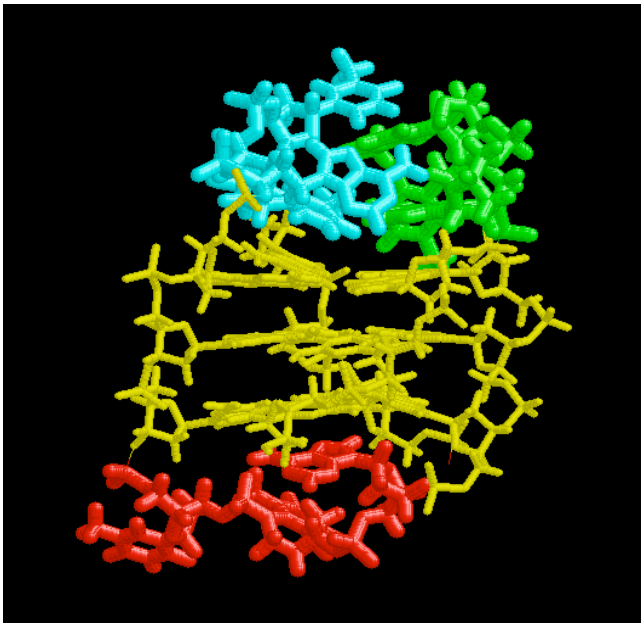


Figure 3.3.8. Wang-Patel antiparallel NMR structure for the human telomeric repeat. G-tetrads are shown in yellow. Loop 1 is cyan, loop 2 is red and loop 3 is green. It can be seen that there are considerable interaction between loops 1 and 3, but loop 2 is remote.

Values for the Pearson correlation and statistical p-value (significance level) are shown below for each pair of loops. The correlations between loops 1 & 2, and between 2 & 3 are not significant. Those between 1 & 3 are significant at $p < 0.005$ and positive, as predicted. Diad windowed simulates of chromosome 21 were also studied and showed no significant correlations between any of the loops, as expected.

Loops	Correlation	Significance	Comments
1 & 2	-0.05	0.152	Not significant (p > 0.005)
1 & 3	0.10	0.004	Significant (p < 0.005)
2 & 3	-0.07	0.039	Not significant (p > 0.005)

Table 3.3.10. Pearson correlations and p-values comparing each loop for relatively long-looped quadruplexes of chromosome 21. Values were calculated using MINITAB. Only the correlation between loops 1 & 3 is found to be statistically significant at the p< 0.005 level.

3.3.6.4 Parallel folds

When a quadruplex has a very short loop, it tends to fold in a parallel form with double chain reversal loops (see Chapter 2). In this form, the loops project in a propeller-like fashion, and do not have the opportunity to interact with each other. This is exemplified by the structure in 3.3.9 for the parallel fold of the human telomeric quadruplex.[29] As a result, no correlations would be predicted between the loop lengths. However, following on from the arguments in chapter 2, it might be predicted that if any of the loops are short and induce a parallel fold, then if there were a positive selection pressure for stable quadruplexes, then the other loops would also be shorter than found otherwise.

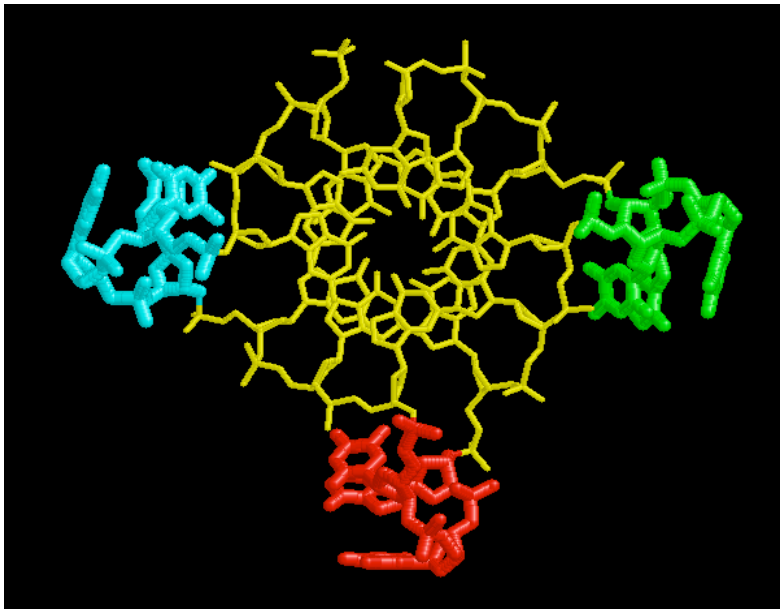


Figure 3.3.9. Neidle parallel crystal structure of the human telomeric repeat. Colours are as in figure 3.3.6. All three loops are remote from each other and cannot interact. K^+ ions are not shown.

To investigate this, the subset of loops where the second loop was short (1 or 2 bases long) was compared to that where the second loop was long (3-7 bases). The lengths of the 1st and 3rd loops were then compared in each case using the one-way ANOVA calculations of MINITAB[169]. Using the 95% confidence interval, it was found that the 1st and 3rd loops were significantly shorter in the subset where the second loop was short than in the subset where the second loop was long. The similarity of the 1st and 3rd loops was

also confirmed, as there is no significant difference between them in either subset.

Loop	Length of loop 2	Mean length of sample loop	95% Confidence range of mean
1	> 2	3.51	3.43 – 3.61
1	< 3	3.29	3.19 – 3.39
3	> 2	3.53	3.45 – 3.63
3	< 3	3.25	3.15 – 3.35

Table 3.3.11. One-way ANOVA results comparing loops 1 and 3 for short (l<3) and long (l>2) lengths of loop 2. Loops 1 and 3 are statistically identical, and significantly shorter at the 95% confidence level in the 'short' subset than the 'long' subset.

3.3.5.4 Whole-genome analysis

In order to study the entire human genome, it was decided to perform a more detailed statistical analysis in order to be able to make comments about the direction and significance of the deviations from chance for the loop interactions. In order to do this, *quadparser* was used to generate a 7x7x7 matrix for all the PQS in the genome, containing in cell {i,j,k} the number of PQS with loop lengths {i,j,k} in the genome. This is represented as a mosaic plot in figure 3.3.10

We initially attempted to represent these data as three independent distributions. However it was immediately clear that there are patterns within this data that deviate from randomness. In particular, it was clear that there is a peak of frequency for values where the length of the first loop is equal to that of the third loop (i=k). As a result, fitting to independent distributions left a 'spine' of excess sequences occurring along this diagonal. This spine requires an extra term, and the whole frequency expression may be expressed mathematically as below, where $N_{\{ik\}}$ is the predicted count with first loop length i and third loop length k, β_i and β_k are the two independent

distributions, *a* is a constant describing the population of the 'spine', and α_i describes the distribution of sequences along the spine.

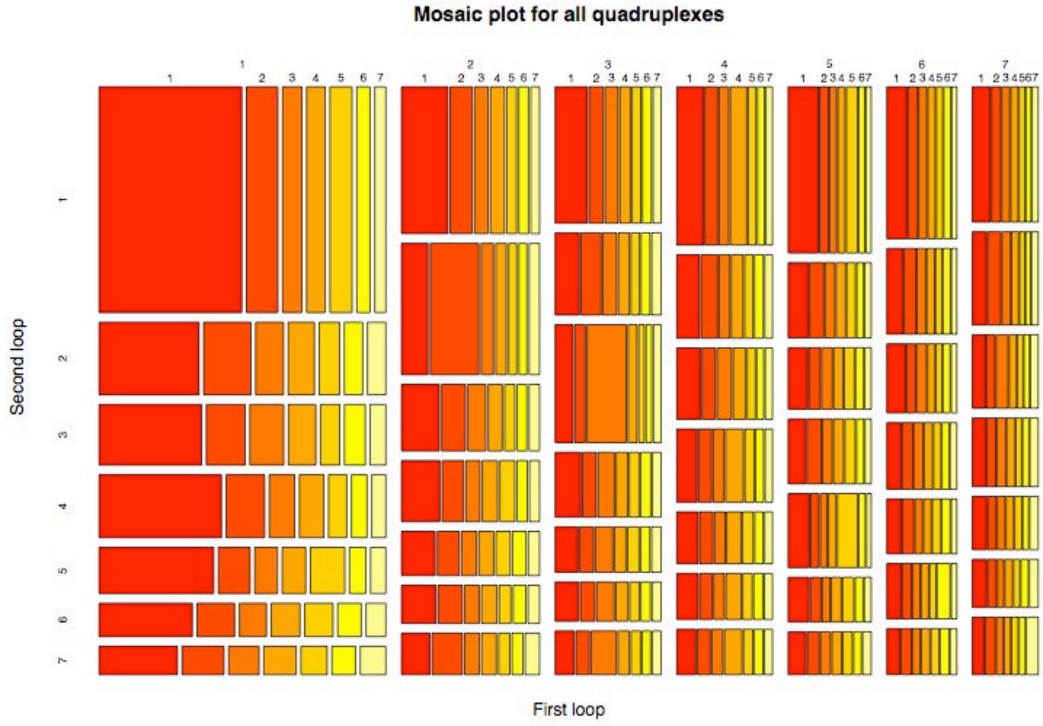


Figure 3.3.10 Mosaic plot representing the loop lengths of all putative quadruplexes found in the human genome. The seven principle columns represent the lengths of the first loop, the seven rows the lengths of the second loop, and the seven segments in each box the lengths of the third loop. The area of each box is proportional to the number of sequences found with that combination of loop lengths. The plot was produced using the program *R*, using the command mosaicplot(…).

$$N_{\{ik\}} = a \cdot \alpha_i + (1-a)\beta_i\beta_k \quad \text{for } i = k$$
$$N_{\{ik\}} = (1-a)\beta_i\beta_k \quad \text{for } i \neq k$$

This is called diagonal quasi-independence,[169] and corresponds to a probability mixture model, in which with a probability *a*, the loops lengths are the same, and with probability (1 - *a*), they are independent. Fitting the data to the model gives a value for *a* of 0.09, implying that approximately 10% of the potential quadruplex patterns show a high correlation between the loop lengths.

These results show that the distribution of frequencies of PQS with various loop lengths is significantly non-random. It would be extremely difficult to rationalise any of these observations without invoking a quadruplex model, and in particular, if these sequences were present as duplexes only, such links between 'loops' would be inexplicable. Hence at least some of the PQS found must form quadruplexes physiologically under some conditions.

What do all of these quadruplexes do? Which of the PQS are actually relevant? In the next sections I investigate the co-incidence of PQS with other biologically relevant entities, single nucleotide polymorphisms (SNPs) and conserved regions, in an effort to identify which are naturally important.

3.4 Cross-genome conservation

3.4.1 Rationale

It is generally accepted that sequence conservation between distantly-related species is supportive of functionality for the conserved region. If there were no functionality, there ought to be no selective pressure against disruptive mutations, whereas if there is important functionality, there would be significant evolutionary pressure against mutation.

Comparing the human genome to other species can then provide some insight into which of the human PQS are important, and which are merely chance occurrences. In this section I discuss evidence showing that at least some PQS are conserved between *Homo sapiens* and the common mouse, *Mus musculus*. I begin with considering one of the sequences studied in chapter 2, the *c-kit* quadruplex, and then address wider questions of conservation of sequences near conserved genes and in a sample chromosome, chromosome 21.

3.4.2 C-kit

A number of genes, especially those involved in the more fundamental processes, have been extensively studied and homologous genes identified in the genomes of other, non-human species. This provides the opportunity to

examine specific gene sequences to see if the conservation in the gene coding regions extends outside the gene itself.

I initially examined the sequences that were studied biophysically in Chapter 2, comparing *Homo sapiens*, *Mus musculus* and the Norwegian rat *Rattus norvegicus*. Data on all three species is available from ENSEMBL, as is information on those genes found to be homologous between them all.

In humans, the *c-kit* proto-oncogene has two putative quadruplex sequences just upstream of the transcription and translation start sites, in the promoter region for the gene. Rats and mice each have one in this region, located in a very similar position and with a similar sequence to one of the human sequences – ***c-kit-2*** in the nomenclature of Chapter 2. Details of the sequences and positions of these PQS, along with an alignment for the homologous sequences are shown below.

Species	Sequence	Distance 5' from start site for:	
		Transcription	Translation
Homo sapiens	AGAGGGAGGGCGCTGGGAGGAGGGGCTG	84	105
Homo sapiens	CCCGGGCGGGCGCGAGGGAGGGGAGG	136	157
Mus musculus	CCCGGGCGGGAGAAGGGAGGGGCGT	98	142
Rattus norvegicus	CCCGGGCGGGAGAAGGGAGGGGCGT	97	141

Table 3.4.1. Locations and sequences of PQS in the 5' upstream regions of genes homologous to the human *c-kit* gene in mice and rats.

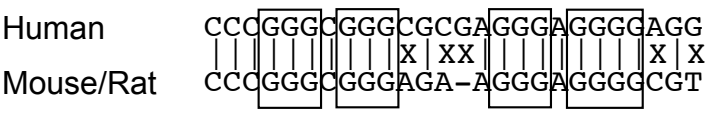


Figure 3.4.1. Alignment of the conserved regions of the c-kit quadruplex between that found in *Homo sapiens* and in *Mus musculus* and *Rattus norvegicus*. These last two are identical. The position of the gap shown for the mouse/rat sequence is arbitrary.

This conservation is very strong, with no differences between the two sequences in the tetrad-forming regions or the sequences that would form

loops one and three. Only the long central loop appears to vary, and it is considerably different in the two cases. Since the conservation of the 5' upstream regions of these genes are only weakly similar, this pocket of conservation suggests that there may be some functionality ascribable to this PQS.

Further evidence for the importance of this region may be found in work by Park *et al.*,[170] who showed that a 41 bp region upstream of the 5' transcription initiation site was crucial for maximal core promoter activity. Interestingly, their 41bp region starts with the **c-kit-2** sequence, providing further evidence that this is an important sequence. They advance the hypothesis that the binding of a transcription factor called Sp1 may be the critical factor to this region, but there is no reason why this could not be in addition the effects of quadruplex formation.

The other sequences discussed in Chapter 2 did not show conservation. In some cases, this is just due to an absence of homologous genes as predicted by ENSEMBL, but may also suggest that these sequences do not play a physiological role sufficiently important to merit conservation. This of course does not preclude them from acting as useful drug targets; indeed, the c-myc target Hurley has studied is not conserved yet still appears functional.

3.4.3 Global search results

In order to develop this approach further, the search was then extended to the rest of the genome. Two approaches were used. The first of these was to examine the 5' regions (5' untranslated region (UTR) and 1000 bases upstream) of genes listed in ENSEMBL as being homologous between the three species examined. Those genes with quadruplexes in each of them were output, and the results (12 genes) are shown in Appendix C. This region was selected as a core region showing promoter activity, and hence any conserved quadruplexes identified would have a higher possibility of being significant.

A few of these appear to be chance events, as the sequences of the quadruplexes are not similar across the species. Most, however, are very well conserved, especially considering that the regions of the genome under consideration are non-coding and hence under lower selective pressure than coding regions. Nonetheless, these results almost certainly underestimate the true number of matches existing, for two reasons. Firstly, we do not yet have a complete list of all homologous genes, so miss a large number of potential matches. Secondly, the limited nature of the regions studied may mean that sequences further than 1000 bases upstream may be missed, as may sequences in exons or introns. The sequences found, however, do provide useful initial candidates for further exploration.

Another approach is to map the location of quadruplexes along a chromosome, and compare the resulting distribution to the locations of conserved regions. Comparing these sequences to the local GC content also gives an idea of whether the quadruplex patterns may be predicted.

This has been performed for chromosome 21, one of the best studied chromosomes. The results show a considerable number of quadruplexes towards the 3' end of the chromosome, in a region known to be GC- and gene-rich (see figure 3.3.4). Less expectedly, there is also a peak just 5' of the centromere in a low-GC area, where a high frequency of conserved elements is also found. It is not clear what the role of this region is – it is relatively sparse in genes, but it must presumably have a function or else it would not be conserved.

3.4.4 Discussion

If quadruplex sequences are involved in physiological function, and have undergone selective pressures in favour of their stability, then it is likely that they would be conserved between species. This would also allow for an appropriate time period in which selection could occur.

One problem is identifying which regions are conserved between different species, and this is very much an unresolved problem. Also, because we do not yet know what these quadruplexes do, it is hard to define exactly where they would be located, or that they would be in the same position relative to a gene in different species.

However, by accepting currently identified homologies and defining a relatively narrow area to look for quadruplexes, a number of conserved quadruplexes have been identified in a key region for gene promotion. These are good candidates for having physiological activity.

3.5 SNPs and quadruplexes

Another important source of genomic information is the ever-growing database of single nucleotide polymorphisms (SNPs). SNPs that occur in physiologically important quadruplexes may destabilise (or overstabilise) the quadruplexes, resulting in malfunction of the genetic controls, which may be observed as a disease or syndrome. Such observations would support the hypothesis that quadruplexes play a biological role, and would allow a deeper understanding of the conditions that result.

There are currently approximately 8.7 million SNPs (including insertion/deletions) in v.121 of the dbSNP database. Comparison of the positional data for these (from the ENSEMBL database using NCBI build 35) with quadruplex positions using the same build of the human genome shows that 21,896 occur within PQS regions. If SNPs were randomly distributed, then approximately 25,000 would be expected – somewhat higher than found. This could potentially be supportive of functionality, as SNPs tend to be relatively rare in functional regions such as coding regions. However, it is very difficult to draw any absolute conclusions, as the quality of the data for SNPs is relatively poor; A large number have only been reported once and may be spurious, and because there have been detailed investigations of particular areas, they are not evenly distributed through the genome.

There is now increasing interest in thorough SNP cataloguing and annotation, and two projects in particular promise to generate important information relating to SNP/function linkage. One is the HapMap project, focused on sequencing SNPs in different populations to look for important SNPs and genes associated with human disease and pharmaceutical response. The other is the ENCODE project, which aims to provide a much higher level of genomic annotation by focusing on 1% of the entire human genome. One of the annotation levels they are investigating is SNP genotyping in considerable depth. Any features arising from either program that are located in regions with a capability of forming quadruplexes will be of great interest, and by using *quadparser* it will be easy to investigate this.

Even in the absence of such broad data sets, other sequences of interest can be examined, and in table 3.5.1 below, I show five examples of SNPs located inside quadruplex forming regions in the 5' UTR and 1000 bp upstream of genes known to be involved with disease.

Protein	**Diseases**	**Sequence**	**SNP**
p45	Breast Cancer Rhabdomyosarcoma	tcaCCCCaccggcaCC**M**gtgCCCaagCCCgCCCCtgc	A/C
PMS1	Colorectal cancer, hereditary	gctGGGtgcGGGtgcGGGtgcGG**K**GttGGGcct	G/T
TRH	Thyrotropin-releasing hormone deficiency	gatCCCgagtCCCCggatCCCgga**S**CCatc	C/G
Secreto-granin III	Muscular dystrophy	gccgCCCagtCCCggCCCCtct**S**CCgCCCCacaCCCaCCCtcc	C/G
Angiotensin-Converting Enzyme	Myocardial infarction susceptibility	acc**GS**GtcatGGGGGccgcctcGGGccgccGGGGGccGGGGctg	C/G

Table 3.5.1. Disease related genes with SNPs in PQS regions which could disrupt quadruplex formation in PQS in the 5' upstream or 5' UTR region. SNP codes are shown using the IUPAC nomenclature.

3.5.2 COL1A1

Osteoporosis is a common disease, especially in the elderly. There are various genes with an important role in regulating this. One of these is the COL1A1 gene, which encodes the α1(I) protein chain of type 1 collagen, the major bone protein. Grant and coworkers[171] have identified a G to T SNP that has been associated with lowered bone mineral density and increased risk of osteoporosis fracture.

The relevant sequence is shown below, and is located in towards the 3' end of the 1st intron of the gene, on the sense strand. The two alleles are (unhelpfully) labelled as 'S' (the predominant form) and 's'.

'S' allele GGGAATG**G**GGGCGGGATGAGGG

's' allele T

According to Mann and coworkers,[172] Ss heterozygotes had approximately 3 times as many primary RNA transcripts from the s allele as the S allele, and Ss individuals had a lower yield strength of bone than SS individuals, due to an excess of α1(I) protein over α2(I).

What is the reason for the overactivity of the 's' allele of the promoter? One explanation is that the zinc-finger transcription factor (TF) Sp1, which has a binding site adjacent to the SNP and acts to increase gene expression, binds more tightly to the 's' allele.[172] An alternative explanation would argue that the quadruplex formed from the 's' allele would be less stable due to the removal of the G residue underlined above. Hence, this strand would spend more time as a duplex than with the 'S' allele, resulting in greater accessibility for transcription.

An alternative hypothesis would combine the Sp1 approach with the quadruplex, suggesting that the 'S' allele, which forms a quadruplex more

readily, would be less available for Sp1 binding that the 's' allele, and so would be less activated.

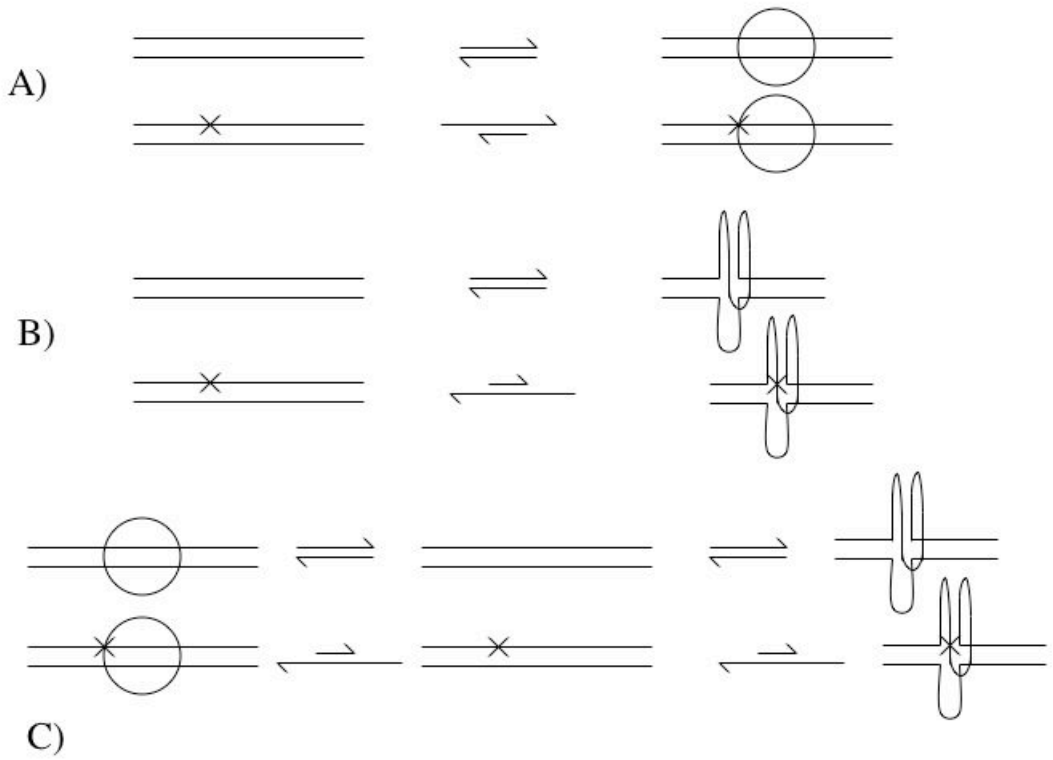


Figure 3.5.1. Three possible explanations for the observed effect of the mutation in COL1A1. A) Sp1 (round circle) binding is preferred for the 's' allele (cross). B) Quadruplex formation is disfavoured for the 's' allele. Since quadruplexes have been shown to block transcription, this reactivates the gene. C) competition equilibrium. The 's' allele disfavours the quadruplex and hence is more available for Sp1 binding.

It is interesting to note that Sp1[173,174] and other similar TFs that bind G-rich domains[175] would be expected to bind many of the regions identified as PQS, and a search of the TRANSFAC database of known TF binding sites[176] reveals many hundreds of known TF binding sites that could also form quadruplexes. This includes a rather extreme category of binding site, the so-called 'G-strings',[177] which consist of a run of 10 or more guanines.

3.6 Conclusions

There are a large number of potential quadruplexes in the human genome, some 350,000 all together. Part of this high frequency is due to the structured

nature of the human DNA, which has a bias to homodiads and exhibits highly variable base composition across genomes. This makes the development of control sequences complicated, and it is not possible to say definitively whether quadruplex-forming sequences are over- or under-represented in the genome.

Examining statistical properties of the loops in the quadruplexes, however, makes it clear that there are selective pressures acting on these sequences in ways which cannot easily be explained without suggesting that many (at least 10%) of the potential quadruplexes do actually exist, and have been subject to selective pressures. The fact that the observed trends correspond to stabilising rather than destabilising interactions suggests that the quadruplexes are selected for, rather than against. This is the first evidence of wide-spread functionality for genomic quadruplexes.

Having proposed that there is a physiological meaning to at least some of the potential quadruplexes, the problem now turns to that of identifying potential candidates for activity. One approach is to examine quadruplex-forming sequences that are conserved across genomes, as these might be predicted to have functionality, which is why they are conserved. This approach reveals a number of candidate genes, including one of those studied biophysically in the previous chapter.

The other method for identifying interesting candidates is to look at single nucleotide polymorphisms (SNPs). There are ongoing projects to map these mutations to disease states and gene expression levels, and any such disease- or expression-correlated SNPs which could form part of a quadruplex immediately provide a candidate sequences and hypothesis to explain the effect of the mutation. COL1A1, a mutation in which can cause osteoporosis, is one such candidate, and many more will arise as the projects continue.

CHAPTER 4

Conclusions and perspectives

When I began my doctoral studies, quadruplex DNA was largely considered as a structure with relatively little attention being paid to its function.[11,12,178] Much of the work that had been performed consisted of structural studies, with only limited physiological applications, most notably chromosomal telomeres.[43,47,51] My own initial work was also directed to structural questions, and the design of new structural motifs.

The other key focus in the field was on telomerase 'inhibition' (more correctly, substrate inactivation), and a variety of small molecule binders were being developed both in our group and others, with a view to downregulating telomerase activity in cancerous cells.[60 68]

Then, in 2002, Hurley published a paper[125] showing that a putative quadruplex in the *c-myc* promoter was connected to transcriptional activity and that its effect could be decreased by preventing quadruplex formation and increased by stabilising it using a small molecule.

I was inspired by this demonstration to address the general question of the role that genomic quadruplexes could play, either as drug targets or as a natural biological control mechanism. There are many possible roles that quadruplexes could play; they could be imagined to mark out transcription factor binding sites, splice sites or to provide a steric block to transcription, as well as many other possibilities.

In order to tackle this genomic question, the first issue to address was how to define a quadruplex from its primary sequence. Chapter 2 presents this work and the arguments leading up to the development of the 'folding rule'. Some aspects of this are based on existing literature, other sections required additional experimentation.

One area that required experimentation was the role of the loops in quadruplex stability. To understand this question better, a model series of oligonucleotides with variable loop lengths was studied using biophysics and

molecular modelling techniques. This gave considerable insight into the stability of quadruplexes with different loop lengths and the manner in which they tend to fold.

There are many factors that affect quadruplex structure and stability, and the folding rule only captures some of them. It certainly excludes a number of sequences which are known to form quadruplexes, such as the thrombin binding aptamer (TBA).[95,96] However, it is at worst an underestimate of how many quadruplexes can form, and to date there have been no reports of sequences which conform to the rule not forming quadruplexes.

Using this rule, it is possible to search large-scale regions of the genome for putative quadruplex sequences (PQS). To test the concept, a handful of genes of therapeutic interest were initially selected, and a large number of potential quadruplexes were identified in each of them. Some of these were selected for biophysical testing, and they all proved to form quadruplexes *in vitro*.

In order to further confirm their structure, and to demonstrate that it would be possible to find compounds that would selectively bind these quadruplexes, we performed binding assays with a selection of compounds, using some of our repertoire of small molecule binders and an engineered three zinc finger quadruplex binding protein. All of the quadruplexes showed good binding properties, with some selectivity.

In order to demonstrate further the potential for these quadruplexes to be relevant therapeutically, studies were performed on a mutant form of the fruit fly *Drosophila melanogaster*, which has a quadruplex-containing promoter fused to a reporter gene. A number of quadruplex ligands were tested on the embryos, but to date gene deactivation, has not been demonstrated.

Chapter 3 describes studies on the entire genome, to understand the prevalence of potential quadruplexes, and whether there is evidence of any selective pressures on them. A computer program was developed, capable of

rapidly identifying potential quadruplexes in large amounts of genomic DNA, and using this was able to count for the first time the number of potential sequences in the genome. The exact value depends on exactly how they are counted, but is around 350,000.

If these are physiologically relevant, then they may have been subject to selective pressures, either favouring their formation or disfavouring it. In order to find out if this had occurred, calculations were performed on how many would be expected in the genome, assuming the genome to be a random sequence of bases. This gave a value that was orders of magnitude too low, both for normal quadruplexes and 'control' AT-quadruplexes, which have no known physical meaning, but are merely a pattern in duplex DNA.

By allowing for some of the structural aspects of genomic DNA, an alternative method of modelling DNA was developed, allowing the creation of simulates of each chromosome that preserved aspects of the structure but were otherwise randomised. This gave a prediction of quadruplex frequency that was much closer to that found, both for the real quadruplexes and the AT-patterns. By adjusting the window size, it was possible to calibrate the model to correctly predict the number of AT-patterns, and hence to propose that GC-patterns were under-represented by around 37%.

It was then considered whether evidence for selective pressures may be found in the loop regions. If the quadruplexes are real, then there may be some non-random structure to the three loops, but if they are not real, and exist as duplex DNA physiologically, then all three loops should be independent. It was shown that the loops were not the same, and that the central loop has a different distribution of lengths to the others. There are also clear correlations between the lengths of loops 1 and 3, something that would be very hard to rationalise without invoking quadruplex formation, or some other secondary structure.

The problem of identifying which quadruplexes are likely to be important was then addressed, and the presence of cross-genome correlation used as an

indication of evolutionary importance. A wide variety of conserved quadruplexes are identified. Single nucleotide polymorphisms (SNPs) provide a useful link between sequences and phenotypic effects. A collaboration with the Sanger Institute and their projects to correlate SNPs with gene activity has been initiated, and some candidates developed where there is a known medical effect related to an SNP. One example related to osteoporosis is presented, along with a mechanistic hypothesis to explain how the phenotypic effect could occur.

This work so far has explored a new field for understanding the role of DNA secondary structure in physiological function and disease. It has already developed a large number of candidate genes for drug targeting, along with proof that drug binding can be achieved. Many more will flow from whole-genome analyses and collaboration with other external projects, such as HapMap and ENCODE.

There are many future directions this project should take, and there are a number of new projects in the Balasubramanian lab, either in progress or awaiting funding, which stem from the work described in this thesis. This includes further work characterising the quadruplex-forming sequences **c-kit-1**, **c-kit-2**, and **N-ras**, and the development of further drugs which bind to them. This will eventually lead to *in vitro* transcription / translation studies on these systems.

One aspect of genomic quadruplex formation which has received relatively little attention (with the honourable exceptions of studies by Fox[54] and Sugimoto[55,56]) is the effect of having a complementary strand present. This will change the equilibrium of the G-rich strand away from quadruplex formation, by forming a duplex. The complementary strand may itself form an *i*-motif structure, also perturbing the equilibrium. This is of particular importance in systems more akin to natural chromosomes, as the complementary strand is not only present in solution, as studied by Fox and Sugimoto, but is attached to the G-rich strand at either end by long stretches

of duplex, resulting in a very high local concentration. Further studies are required to elucidate the details of duplex/quadruplex dynamics.

Because of the effect of the complementary strand, quadruplexes may be more likely to form in single-stranded nucleic acid sequences. One important occasion when this occurs is in RNA, and especially mRNA. Many examples of putative quadruplexes in mRNA have been identified using *quadparser*, and a project has recently begun to investigate their properties, in particular looking at sequences forming from **N-ras** and the oestrogen receptor. Another series of locations that are of interest are the nuclease hypersensitive sites, such as in *c-myc*.[120,121] Part of the ENCODE project will include producing lists of these sites, which can then be processed with *quadparser* to look for potential quadruplex sequences.

The development of ligands that are selective for only one type of quadruplex is a real challenge, and may lead eventually to candidates for therapeutic testing. Successful demonstration of *in vivo* activity is also important, either in the *Drosophila* system described or, if that is unsuccessful, a simpler system, such as bacteria or mammalian tissue culture /cell biological studies.

When the ENCODE data from the Sanger Institute, connecting SNPs to variations in RNA expression level, is available, it will be possible to list all SNPs in quadruplex-forming regions with correlations to gene expression levels. I believe this project will discover a large number of quadruplexes for which an effect can be identified, which will provide important candidates for study. Similarly, the Cancer Genome Project at the Sanger Institute will generate mutations linked to cancer, and we are discussing with them the possibility of specifically looking at quadruplex-forming regions in this project.

In summary, my work has addressed a new dimension of the quadruplex field, and one that has the potential to open up possible new approaches to therapeutic treatments for conditions such as cancer, as well as advancing our understanding of the mechanisms of genetic regulation.

EXPERIMENTAL

5.1 Oligonucleotides

Oligonucleotides were purchased from Invitrogen (Paisley, UK) and were generally synthesised on a 50 nmol scale and supplied lyophilised. They were made up to 50 μM stock solutions in milliQ water and stored at –20 °C until required.

5.2 Ultraviolet melting

UV melting curves were collected using a Varian Cary 1E UV-visible spectrophotometer, measuring the spectral absorbance at 295 nm, which has been previously identified as a key absorbance region for quadruplex formation.[129] Samples were typically prepared at 2 μM in a buffer containing 100 mM KCl and 10 mM Tris.HCl pH 7.4. These samples were then heat-annealed to 90 °C and then allowed to slow-cool to 4 °C over a period of several hours. In a typical experiment 400 μl of a sample was degassed in a Speedvac for 3 minutes, transferred to a 1 cm path length quartz cuvette and then covered with a layer of mineral oil (Sigma-Aldrich). They were then transferred to the spectrometer and equilibrated at 10 °C for 10 minutes. They were then heated to 90 °C and cooled to 10 °C twice consecutively at 0.1 °C/min, with data collection occurring every 0.5 min on both annealing and melting steps. Blank samples containing only pre-annealed buffer were treated in the same manner and subtracted from the collected data.

Data analysis. T_m values for the sequences were in general determined from van't Hoff analysis of the melting profiles using a custom-written procedure. This analysis assumes a simple two-state equilibrium between the folded and unfolded forms, with linear variations of absorbance with temperature for both species.

The details of the analysis are as follows. Initially, a thermal melting plot is produced which has the typical appearance as figure 5.2.1. It has a linear portion at low temperature, which corresponds to the absorption of a

quadruplex varying with temperature. This is then followed by a sigmoidal region with a general decrease in absorption over increasing temperature; this region corresponds to a melting transition from quadruplex to ssDNA. Finally at high temperatures there is another linear region, reflecting the increasing absorption of ssDNA with temperature.

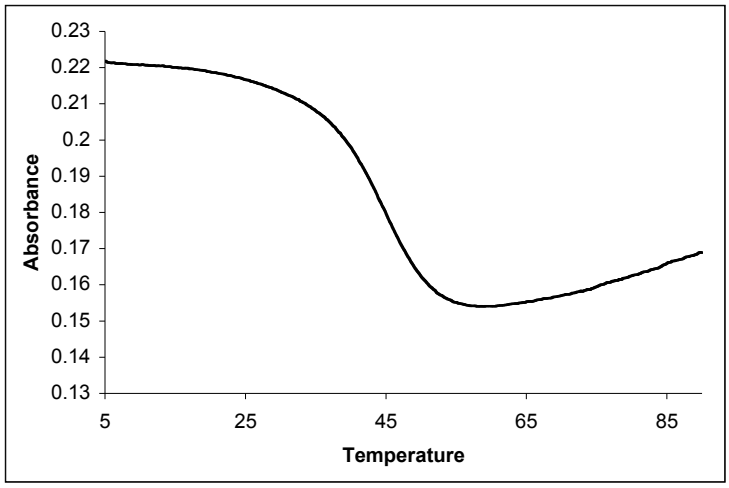


Figure 5.2.1 A typical UV melting profile. This example and those below were prepared using the **6-4** sequence d(TGGGTTTTTTGGGTTTGGGTTTTGGGT).

The first stage in the analysis is to convert the absorbance-temperature plot into a plot of the fraction folded against temperature, by correcting for the linear baselines at high and low temperature. The first step is to fit the high- and low-temperature regions to lines, with a known equation. These may then be used to calculate the fraction folded, α, in terms of the absorbance, A_T, and the extrapolated absorbances based on the linear regions at high and low temperature. This process is shown in figure 5.2.2, and may be expressed mathematically as shown below. The results may be represented as an alpha-plot, an example of which is given as figure 5.2.3

$$\alpha = \frac{A_T - A_{high}}{A_{low} - A_{high}} = \frac{A_T - m_{high} \cdot T - c_{high}}{(m_{low} - m_{high}) \cdot T + (c_{low} - c_{high})}$$

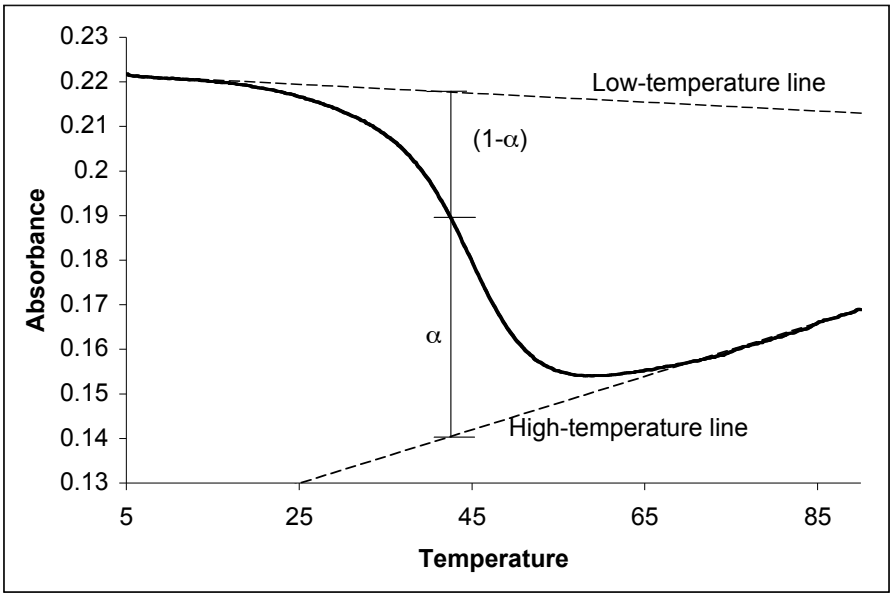


Figure 5.2.2. Analysis of a UV melting profile. Linear fits to the high- and low-temperature regions are performed, and the proportion folded at any given temperature calculated by comparing the measured absorption to the extrapolated linear regions.

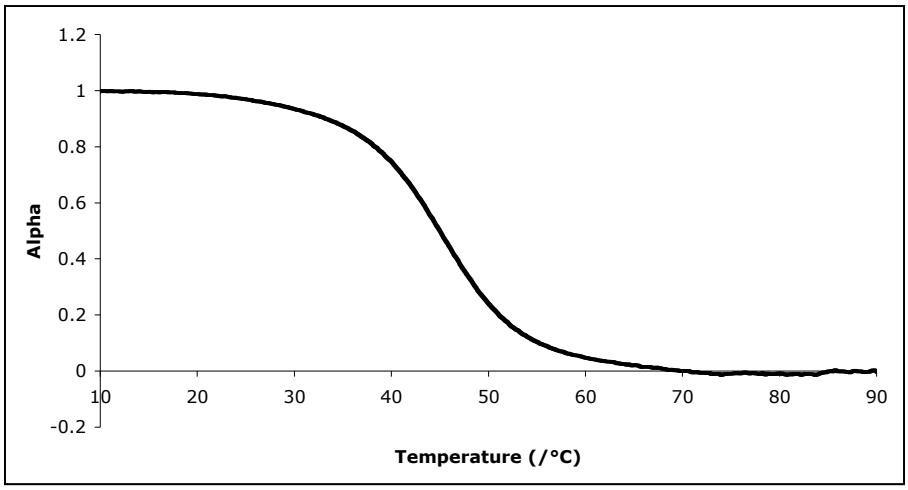


Figure 5.2.3 Alpha curve resulting from melting profile analysis. An alpha curve describes the fraction folded at any temperature, and takes values from 0 to 1. The T_m is where $\alpha = 0.5$ exactly.

Assuming that the quadruplex concerned is unimolecular, the equilibrium constant K_{eq} describing the melting may be expressed simply in terms of α and the total concentration of DNA, C_T, a term which cancels out of the expression.

$$K_{eq} = \frac{[Quad]}{[ssDNA]} = \frac{\alpha \cdot C_T}{(1-\alpha)\cdot C_T} = \alpha \big/ (1-\alpha)$$

Using the known relationship between ΔG and K_{eq}, this expression can be recast to give a linear relationship between $\ln K_{eq}$ and 1/T, with a slope of –ΔH/R and an intercept of ΔS/R. This allows the calculation of T_m = ΔH/ΔS. A sample plot of $\ln K_{eq}$ versus 1/T is shown as figure 5.2.4.

$$\Delta G = -RT \ln K_{eq} = \Delta H - T\Delta S$$

$$\ln K_{eq} = -\frac{\Delta H}{R} \cdot \frac{1}{T} + \frac{\Delta S}{R}$$

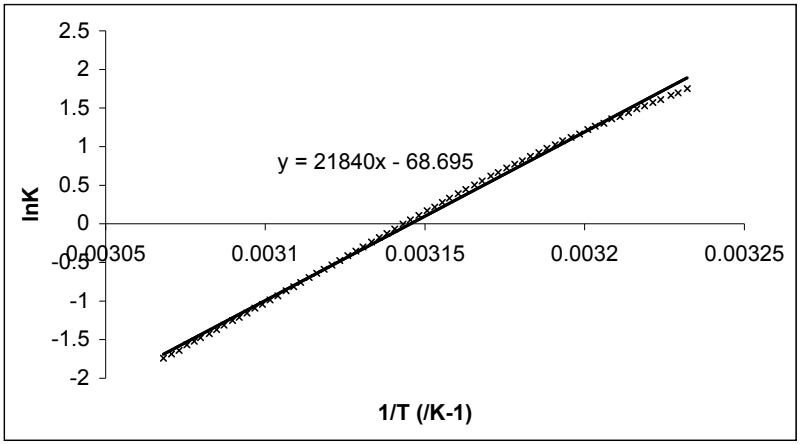


Figure 5.2.4. Van't Hoff plot showing the linear relationship between lnK and 1/T. The slope is –ΔH/R and the intercept ΔS/R

Although the T_m values were highly reproducible, the ΔH and ΔS values themselves are not reproducible, and for that reason are not listed in detail throughout this thesis. One reason for this is the covariation of ΔH and ΔS, and the other is that the ΔH value is only meaningful in this context if the change in heat capacity, ΔC_p, is 0. More accurate results may be obtained by isothermal calorimetry. Typical results for ΔH and ΔS for the systems studied in this thesis were ΔH = -150 kJ mol^{-1} and ΔS = -460 $JK^{-1}mol^{-1}$.

5.3 Circular dichroism

Circular dichroism experiments were performed on a Jasco J-810 spectropolarimeter using inbuilt software. Samples were typically prepared at 4 μM in a buffer containing 100 mM KCl and 10mM Tris.HCl pH 7.4. These samples were then heat-annealed to 90 °C and then allowed to slow-cool to 4 °C over a period of several hours. In a typical experiment 400 μl of a sample was transferred to a 1 cm path length quartz cuvette, and scans performed over the range 220-320 nm. 220 nm is the lowest wavelength achievable, because Tris absorbs strongly below 220 nm, leading to a rise in photomultiplier voltage and unreliable results. Each trace is the result of the average of five scans at 50 nm/min, with a 2 s response time, 1 nm pitch and 1 nm bandwidth. Samples were left to equilibrate at 4 °C for 10 min before scans were performed. A constant flow of dry nitrogen ensured there was no condensation. A blank sample containing only buffer was treated in the same manner and subtracted from the collected data. For graphing purposes, the data was smoothed slightly, with each data point being graphed as the average of those around it and a zero-correction being applied at 320 nm.

5.4 Molecular dynamics

Three template quadruplex structures were obtained from the Protein Data Bank for use in the simulations. The parallel conformation was based on the human telomeric repeat d[$AG_3(T_2AG_3)_3$] crystal structure (PDB code 1KF1). This structure contains three T_2A strand-reversal loops. The NMR structure of the above human sequence is antiparallel and contains two lateral loops and a central diagonal loop. The first structure of the PDB entry 143D was used as an antiparallel template. Both structures were used directly after modification of the central loop to the desired length, ranging from T to T_6. For short loop lengths of one or two nucleotides, an alternative antiparallel template with a lateral central loop was derived from the NMR structure of $d(T_2G_4)_4$ (PDB code 186D). This sequence differs from the human sequence by one G to A mutation in each repeat. The 186D structure contains mixture of parallel and antiparallel guanine strands, with two lateral loops and a third strand-reversal

loop. The coordinates of the first PDB entry were used. The G-quartets were left unchanged, the first and third loops were modified to T_2A, and the central loop modified as previously.

All structure manipulations were carried out within the InsightII suite of programs. Suitable loop conformations were generated using a simulated annealing protocol within the Discover3 program. The quartet stem was held fixed while the loops were heated to 1000 K. Molecular dynamics was carried out at this temperature for 2 ps, before slowly cooling to 300 K over 1 ps, and minimising for 10000 steps. The simulated annealing procedure was repeated on average 100 times for each quadruplex. The resulting structures were divided into clusters of similar conformations, and the lowest energy structure of the most populated cluster was typically used as a starting structure. As loop lengths were increased, more and more clusters were obtained with very few structures in each. In those cases, initial structures were selected depending on the energy as well as the presence of favourable interactions within the loops. On average two starting structures were selected for each quadruplex, and only the most stable conformations have been included in the discussion below. This procedure yielded a structure for the parallel T_5 quadruplex with the loop residues facing the guanine quartets, with hydrogen bonding potential between the loop and guanine residues. Structures for the T_4 and T_6 loops were generated from the T_5 loop by respective deletion / addition of a residue and simple vacuum minimisation with the Discover3 program.

Molecular dynamics simulations were carried out on all the selected structures using the Amber 7 package, with the Amber 99 force field. Two K^+ ions were placed between the quartet planes in the parallel quadruplexes, in accord with the x-ray structure. Due to the unavailability of ion location information in the antiparallel quadruplex structures, two K^+ ions were placed between the quartet planes, based on the parallel quadruplex crystal structure. In the 143D antiparallel structures, a third K^+ ion was placed below the G-quartets, within the lateral loop region, as there was sufficient space to accommodate it. The leap module of Amber was used to add enough K^+ ions to neutralise the

systems, which were then solvated by a box of TIP3P water molecules, extending 10 Å from the solute in each direction. The equilibration procedure consisted of 10 steps, beginning with a 1000 step minimisation of the solvent. 25 ps of dynamics of the solvent then enabled the water molecules to fill any gaps between the solvent and solute. The quadruplexes were minimised for 1000 steps, followed by 3 ps of dynamics with a constraint of 25.0 kcal.mol^{-1}. A series of 5 minimisations were carried out, during which the constraint on the DNA was lowered each time by 5.0 kcal.mol^{-1} until it reached zero. After minimisation, the systems were heated to 300 K over 20 ps. MD simulations were carried out at 300 K with a 2 fs time step and SHAKE applied to constrain the bonds containing hydrogen. The Particle Mesh Ewald (PME) method of calculating long-range electrostatic interactions was employed, with a cutoff of 10 Å. The Carnal and Ptraj modules of Amber were used to analyse the simulations.

5.5 Free energy calculations

Free energy calculations were carried out using the MM-PBSA post-processing method in Amber. Briefly, configurations were sampled from MD simulations in explicit solvent. The free energy was determined from molecular mechanics gas phase energies, solvation free energies using continuum solvent methods, and entropic contributions. Additivity of all the free energy components was assumed. The gas phase energies were calculated using the Anal module of Amber. The Delphi program was used to calculate the electrostatic contribution to the solvation free energy. Dielectric constants of 1.0 and 80.0 were used for the solute and solvent respectively. A grid spacing of 0.5 Å was chosen, with the longest linear dimension of the molecule occupying 80 % of the grid. The non-polar contribution to the solvation free energy ($\Delta G_{nonpolar} = \lambda a + b$) was determined using the MSMS program, where a is the surface area, $\lambda = 0.00542$ and b = 0.92 kcal/mol. Snapshots were collected every 20 ps after the equilibration period, giving 150 snapshots over 1 to 4 ns of each trajectory. As previously shown,[146] the inclusion of the channel K^+ ions is important in free energy calculations, and

three K^+ ions were therefore included in all the calculations. Where only two ions were present within the core of the quadruplex, the closest K^+ ion to the solute was included. The continuum model of electrostatic interactions implicitly averages over the solvent degrees of freedom, but the entropy of the solute is not included and must be calculated separately. This is more computationally demanding than the enthalpy calculations, and was therefore carried out only on a subset of structures taken every 200 ps. The Nmode module of Amber was used to calculate the translational, rotational and vibrational partition functions after minimisation of the structures to a 10^{-5} root mean square gradient. A distance dependent dielectric function was used to mimic the effect of the solvent during the minimisations (dielectric constant of 4). The minimisations caused some structures to become slightly distorted, and these were not included in the entropy calculations (although this did not have an important effect on the entropies calculated).

APPENDICES

Appendix A

Analytical solution for the frequency of PQS in genomic DNA

In this appendix, I calculate how the expected density of quadruplex-forming sequences in genomic DNA varies with varying guanine frequencies, based on the following assumptions:

- DNA may be described as a simple Bernoulli sequence; with each position being independent of all others.
- The length of DNA being examined is very much longer than the maximum length of a quadruplex.
- Overlapping quadruplexes are a relatively rare event, and may be neglected.

A PQS is defined as a sequence conforming to the following rule:

$$PQS \equiv G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+}N_{1-7}G_{3+},$$

where G represents a guanine base, N represents any base.
$\overline{G}$ will be used to represent a base other than G.
Consider any position k along the sequence being examined. The probability that k is a starting point for a PQS may be separated into terms as follows:

$$\begin{aligned} P(k \text{ is start}) &= P(\text{appropriate GGG repeats}) \cdot P(\text{appropriate loops}) \\ &= P(\text{3 G bases})^4 \cdot P(\text{one appropriate loop})^3 \\ &= \alpha_{(p)}^3 \cdot p^{12}, \end{aligned}$$

where $\alpha_{(p)}$ describes the frequency of appropriate loops
and p is the probability of an individual base being G.
Now, $\alpha_{(p)} = \sum_{i=1}^{7} S_i(p),$
where $S_i(p)$ is the probability of finding loops of exactly length i. $\alpha_{(p)} > 1,$ because the flexibility in the loop length results in more 'hit' sequences.
The $S_i(p)$ may be expressed as follows.

$$S_i(p) \equiv \begin{cases} 1, & (i = 1) \\ P(\text{First } i \text{ bases have no } NGGG \text{ sequence, and end in } \overline{G}) & (i \neq 1) \end{cases}$$

The reason for this is to avoid overcounting - if the last base is a *G*, the sequence has alrea been counted with a loop length of *i* -1. Similarly, if there is a *GGG* sequence within the 'loo it has already been counted. If *i* = 1, then any base is acceptable - a sequence of the form (*GGG*)*G*(*GGG*) is considered as two *GGG* runs with a 1-base loop.

So, considering each length in turn:

Value	allowed sequences
$S_1 = 1$	N
$S_2 = (1-p)$	$N\overline{G}$
$S_3 = (1-p)$	$NN\overline{G}$
$S_4 = (1-p)$	$NNN\overline{G}$
$S_5 = (1-p)(1-p^3)$	$N\left\{NN\overline{G}\right\}\overline{G}$
$S_6 = (1-p)\left[1-p^2+p^2(1-p)^2\right]$	$NN\left\{N\overline{G}\right\}N\overline{G}$ *or* $N\overline{G}G\overline{G}G\overline{G}$
$S_7 = (1-p)\left[1-p+p(1-p)^2+2p^2(1-p)^2\right]$	$NNN\overline{G}NN\overline{G}$, $NN\overline{G}G\overline{G}N\overline{G}$, $NN\overline{G}G\overline{G}G\overline{G}$ *or* $N\overline{G}G\overline{G}G\overline{G}N$

So, $\alpha_{(p)} = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7$

$= 7 - 6p - 6p^3 + 9p^4 - 3p^5$

PQS density $\rho(PQS) = P$(k is a start point)

$= \alpha_{(p)}^3 \cdot p^{12}$

$= p^{12} \cdot \left[7 - 6p - 6p^3 + 9p^4 - 3p^5\right]^3$

So, $\boxed{\rho(PQS) = 343p^{12} - 882p^{13} + 756p^{14} - 1098p^{15} + 2835p^{16} - 3357p^{17} + 2484^{18} \ldots}$

For relatively small *p*, small orders of *p* dominate.

Appendix B

Quadparser: a program for rapid searching of DNA for putative quadruplex sequences

By Julian Huppert, Balasubramanian group, Cambridge University Chemical Laboratories and Simon Rodgers, thaze.com

usage:

```
./quadparser <filename> <bases> <# bases in repeat>
<repeats in sequence> <min gap size> <max gap size>
<output file>
```

example: `./quadparser bases.txt CG 3 4 1 7 output.txt`

scans through bases.txt for CCC...CCC...CCC...CCC or GGG...GGG...GGG...GGG, where ... = 1-7 chars, writing into output.txt

optional commands:

-help	-h	gets help
-man	-m	opens the manual
-version	-v	gives version information
-normal	-n	uses standard parameters GC 3 4 1 7 Must go first!

output styles:

-default	-d	default output. Not required
-pos	-p	outputs starting positions only
-coord	-c	outputs first and last positions only. optional prefix as last entry
-number		outputs number of quadruplexes only
-header	-h	outputs hit headers and sequences
-gene	-g	outputs gene name from header only
-loop	-l	outputs table of loop length matches
-ll		abbreviated loop output
-l2		uncollated loop output

Examples:

```
./quadparser -n bases.txt output.txt
./quadparser -n -c bases.txt output.txt 3
./quadparser -pos bases.txt GC 3 4 1 7 output.txt
```

Full code is shown on the following pages.

```
#include <stdio.h>
#include <stdlib.h>
#include <string.h>
#include <signal.h>

#define  VERSION "1.0"
#define  CHROMOSOMESTARTCOUNTMARKER "bases "

char* szFilename = NULL;
char* szBases = NULL;
int nBasesInRepeat = 0;
int nRepeatsInSequence = 0;
int nMinGap = 0;
int nMaxGap = 0;
FILE* fpOutput = NULL;
char* szFileContents = NULL;
char* szFileOutput = NULL;
int nRuns = 0;
int bToFile = 1;
char* szOutputStyle = NULL;
char* szPrefix = NULL;

char *eliminateChar(char* szString, char cToRemove)
{
   char* pPos=szString;
   int nPos=0;
   while(*pPos)
   {
      if (*pPos!=cToRemove)
      {
         szString[nPos]=*pPos;
         nPos++;
      }
      pPos++;
   }
   szString[nPos]=0;
   return szString;
}

long fileLength(FILE* pStream)
{
    if (pStream)
    {
        long lnLength = 0L;
        long lnPos = 0L;
        lnPos = ftell(pStream);
        fseek (pStream, 0, SEEK_END);
        lnLength = ftell(pStream);
        fseek(pStream, lnPos, SEEK_SET);

        return lnLength;
    }
    return 0L;
}

/* Read a file into a buffer just the right size */
long fileReadInToBuffer(FILE* pFile,char** pacBuffer)
{
    /* Make it one bigger to work with text files into a nulled buffer - that way they
are zero terminated too!*/
    long lnBufferLength = fileLength(pFile)+2;
    long lnBytesRead = 0L;
    if (pacBuffer)
    {
        *pacBuffer = (char *)calloc(sizeof(char)*(lnBufferLength),sizeof(char));
        if (*pacBuffer)
        {
         //make sure that you open the file in binary, opening it in ASCII will
convert the carriage
         //returns and lnBytesRead !=(lnBufferLength-2)
            lnBytesRead = fread( *pacBuffer, sizeof(char), lnBufferLength, pFile);
        }
    }
    if(lnBytesRead==0)
   {
      free(*pacBuffer);
   }
    /* Return the actual amount read - not the buffer size */
```

```
   return lnBytesRead;
}

void init(int argc, char *argv[])
{
   FILE* fpInput = NULL;
   int ii,jj,kk;

   if (argc == 2 && strcmp(argv[1], "-version") == 0 || argc==2 && strcmp(argv[1], "-
v") == 0)
   {
      printf("\nquadparser version %s\n", VERSION);
      printf("version history:\n");
      printf("0.1\t04-09-03\tInitial version\n");
      printf("0.2\t11-09-03\tRemoves line feeds\n");
      printf("\t\t\tAccepts FASTA format\n");
      printf("0.3\t16-09-03\tAccepts multiple bases at input\n");
      printf("\t\t\tExports output to file\n");
      printf("\t\t\tChromosome name/number prepends output line\n");
      printf("\t\t\tVersion information\n");
      printf("\t\t\tProper treatment of poly {G} repeats\n");
      printf("0.4\t04-12-03\tFixed detection bug\n");
      printf("\t\t\tTreats CCCCCCC type sequences correctly as (CCC)C(CCC)\n");
      printf("0.5\t11-12-03\tSearches through many concatenated files\n");
      printf("\t\t\tOutputs all information header\n");
      printf("1.0\t26-06-04\tRelease version\n");
      printf("\t\t\tOutput format controllable from input\n");
      printf("\t\t\tLoop counter installed\n");
      printf("\t\t\tGene stripper installed\n");

      printf("\n");
      exit(0);
   }

   if (strcmp(argv[1], "-help") == 0 || strcmp(argv[1], "-h") == 0 || strcmp(argv[1],
"-man") == 0 || strcmp(argv[1], "-m") == 0)
   {
      printf("\n\nquadparser version %S\n", VERSION);
      printf("Programme to search for quadruplex-forming sequences in DNA\n");
      printf("By Julian Huppert, Balasubramanian group, Cambridge University Chemical
Laboratories\n");
      printf("And Simon Rodgers, thaze.com\n\n");
      printf("usage:\n\t%s <filename> <bases> <# bases in repeat> <repeats in
sequence> <min gap size> <max gap size> <output file>\n\n", argv[0]);
      printf("ex:\t %s bases.txt CG 3 4 1 7 output.txt\n", argv[0]);
      printf("\t scans through bases.txt for CCC...CCC...CCC...CCC or
GGG...GGG...GGG...GGG, where ... = 1-7 chars, writing into output.txt\n\n");

      printf("optional commands:\n-help\t-h\tgets help\n-man\t-m\topens the manual\n-
version-v\tgives version information\n");
      printf("-normal\t-n\tuses standard parameters GC 3 4 1 7\tMust go first!\n");
      printf("\noutput styles:\n-default-d\tdefault output. Not required\n");
      printf("-pos\t-p\toutputs starting positions only\n");
      printf("-coord\t-c\toutputs first and last positions only. optional prefix as
last entry\n");
      printf("-number\t\toutputs number of quadruplexes only\n");
      printf("-header\t-h\toutputs hit headers and sequences\n");
      printf("-gene\t-g\toutputs gene name from header only\n");
      printf("-loop\t-l\toutputs table of loop length matches\n");
      printf("-l1\t\tabbreviated loop output\n");
      printf("-l2\t\tuncollated loop output\n");
      printf("\nExamples:\t%s -n bases.txt output.txt\n", argv[0]);
      printf("\t\t%s -n -c bases.txt output.txt 3\n", argv[0]);
      printf("\t\t%s -pos bases.txt GC 3 4 1 7 output.txt\n\n", argv[0]);

      exit(0);
   }

   // allow default setting of parameters CG 3 4 1 7
   if (strcmp(argv[1], "-n") == 0 || strcmp(argv[1], "-normal") == 0)
   {
      szBases = strdup("GC");
      nBasesInRepeat = 3;
      nRepeatsInSequence = 4;
      nMinGap = 1;
      nMaxGap = 7;

      if(argv[2][0] != '-')
      {
```

```
         szOutputStyle = strdup("-d");
         szFilename = strdup(argv[2]);
         szFileOutput = strdup(argv[3]);

      }
      //eg quadparser -n bases.txt output.txt
      else
      {
         szOutputStyle = strdup(argv[2]);
         szFilename = strdup(argv[3]);
         szFileOutput = strdup(argv[4]);

         if(argc>5 && (strcmp(szOutputStyle, "-coord") == 0 || strcmp(szOutputStyle,
"-c") == 0))
            szPrefix=strdup(argv[5]);
      }
      //eg quadparser -n -pos bases.txt output.txt
   }

   else
   {
      if(argv[1][0] != '-')
      {
         szOutputStyle = strdup("-d");
         szFilename = strdup(argv[1]);
         szBases = strdup(argv[2]);
         nBasesInRepeat = atoi(argv[3]);
         nRepeatsInSequence = atoi(argv[4]);
         nMinGap = atoi(argv[5]);
         nMaxGap = atoi(argv[6]);
         szFileOutput = strdup(argv[7]);

      }

      else
      {
         szOutputStyle = strdup(argv[1]);
         szFilename = strdup(argv[2]);
         szBases = strdup(argv[3]);
         nBasesInRepeat = atoi(argv[4]);
         nRepeatsInSequence = atoi(argv[5]);
         nMinGap = atoi(argv[6]);
         nMaxGap = atoi(argv[7]);
         szFileOutput = strdup(argv[8]);

         if(argc >9 && (strcmp(szOutputStyle, "-coord") == 0 || strcmp(szOutputStyle,
"-c") == 0))
            szPrefix=strdup(argv[9]);
      }
   }

   if (!szFilename || !szBases || !nBasesInRepeat || !nRepeatsInSequence || !nMinGap
|| !nMaxGap || !szFileOutput)
   {
      printf("invalid parameters\n");
      exit(0);
   }

// printf("%s   %s   %s   %d   %d   %d   %d   %s\n", szOutputStyle, szFilename,
szBases, nBasesInRepeat, nRepeatsInSequence, nMinGap, nMaxGap, szFileOutput);

   fpInput = fopen(szFilename, "r+b");
   if (!fpInput)
   {
      printf("file %s not found\n", szFilename);
      exit(0);
   }
   fileReadInToBuffer(fpInput, &szFileContents);
   fclose(fpInput);

   if (strcmp(szFileOutput, "stdout") == 0)
   {
      bToFile = 0;
   }
   else
   {
      printf("output to file %s\n", szFileOutput);
      fpOutput = fopen(szFileOutput, "w+b");
```

```
      if (!fpOutput)
      {
         printf("could not open file %s\n", szFileOutput);
            exit(0);

      }
   }
}

int main(int argc, char *argv[])
{
   init(argc, argv);

   if (szFileContents)
   {
      int nBases = strlen(szBases);
      int jj=0;

      char* pPos2 = szFileContents;
      int nBlocks = 0;
      int nSpace = 4;
      char** aszBlocks = calloc(sizeof(char*), nSpace);
      printf("\n");

      // build up an array of block start positions, and blank the kets
      while (1)
      {
         char* pKet = strchr(pPos2, '>');

         if (!pKet || !*pKet || !*(pKet+1))
            break;

         if (nBlocks >= nSpace)
         {
            nSpace *= 2;
            aszBlocks = realloc(aszBlocks, sizeof(char*) * nSpace);
         }

         *pKet = 0;
         pPos2 = pKet+1;

         aszBlocks[nBlocks] = pKet+1;

         nBlocks++;
      }

      for(; jj<nBlocks; jj++)
      {
         char* pStartHeader = aszBlocks[jj];
         int nBase=0;
         int nChromosomeOffset = 0;
         char* pStartSequence = strchr(pStartHeader, '\n');
         char* pEndHeader = NULL;

         if(strcmp(szOutputStyle, "-gene") == 0 || strcmp(szOutputStyle, "-g") == 0)
         {
            pStartHeader=strchr(aszBlocks[jj], '|');
            pStartHeader++;
            pEndHeader = strchr(pStartHeader, '.');
            *pEndHeader = 0;
         }

         if (pStartSequence && *pStartSequence && *(pStartSequence+1))
         {
            char* pChromosomeOffset = NULL;

            *pStartSequence = 0;
            pStartSequence++;

            pChromosomeOffset = strstr(pStartHeader, CHROMOSOMESTARTCOUNTMARKER);
            if (pChromosomeOffset)
               nChromosomeOffset = atoi(pChromosomeOffset +
strlen(CHROMOSOMESTARTCOUNTMARKER));

            //print header if wanted
            if (strcmp(szOutputStyle, "-d") == 0 ||
               strcmp(szOutputStyle, "-default") == 0
               )
```

```
            {
               if (bToFile)
                  fprintf(fpOutput, "%s\n", pStartHeader);
               else
                  printf("%s\n", pStartHeader);
            }

            for (; nBase<nBases; nBase++)
            {
               char* pPos = pStartSequence;
               char* szMatch = calloc(1, nBasesInRepeat + 1);
               char* pStartPos = NULL;
               char* pLastEndSeq = NULL;
               int nSequencesFound = 0;
               int nOverlappingSequencesFound = 0;
               int nRunsFound = 0;
               int bContinue = 1;
               char* pLastPos = NULL;
               char cBase = szBases[nBase];
               int anLoopCount[20][20][20];
               int ii,jj,kk;

               for (ii=0; ii<nMaxGap; ii++)
               {
                  for (jj=0; jj<nMaxGap; jj++)
                  {
                     for (kk=0; kk<nMaxGap; kk++)
                     {

                        anLoopCount[ii][jj][kk]=0;
                     }
                  }
               }

               memset(szMatch, (int)cBase, nBasesInRepeat);

               eliminateChar(pPos, '\r');
               eliminateChar(pPos, '\n');

               //print info titles if writing to screen or if default output style
//             if (bToFile && (strcmp(szOutputStyle, "-d") == 0 ||
strcmp(szOutputStyle, "-default") ==0))
//                fprintf(fpOutput, "\nSearching for %d or more sequences of %d '%c'
bases\n\n", nRepeatsInSequence, nBasesInRepeat, cBase);
//             else if (!bToFile)
//                printf("\nSearching for %d or more sequences of %d '%c' bases\n\n",
nRepeatsInSequence, nBasesInRepeat, cBase);


               while (bContinue)
               {

                  char cTemp = 0;
                  int anRollingLoop [3] = {0,0,0};
                  char* apStartPos [4] = {NULL,NULL,NULL,NULL};
                  ii=0;
                  jj=0;

                  while (1)
                  {
                     int n=0;
                     // find next occurrence of the CCC style sequence
                     pLastPos = pPos;
                     pPos = strstr(pPos, szMatch);

                     // none left - start the next base
                     if (!pPos)
                     {
                        pPos = pLastPos;
                        bContinue = 0;
                        fprintf(fpOutput, "BreakBreakBreak\n");
                        break;
                     }

                     if (ii > 0)
                     {
                        // if this isn't the first sequence in this run, check it's
within nMinGap and nMaxGap from the last one
                        int nGap = pPos - pLastEndSeq;
```

```
                        if (nGap > nMaxGap )
                        {
                           // want to use the last position before the gap got too
long
                           pPos = pLastPos - nMinGap;

                           // move along until we encounter a non-cBase char (may have
long sequence at the end)
                           while (*pPos == cBase)
                              pPos++;

                           break;
                        }
                        else
                        {
                           //shunt rolling loop count along
                           anRollingLoop[0]=anRollingLoop[1];
                           anRollingLoop[1]=anRollingLoop[2];
                           anRollingLoop[2]=nGap;

                           //roll array of Startposes
                           apStartPos[0]=apStartPos[1];
                           apStartPos[1]=apStartPos[2];
                           apStartPos[2]=apStartPos[3];
                           apStartPos[3]=pPos;

                           //if we have three loop values, increment the relevant bit
of the array
                           if(anRollingLoop[0] >0)
                           {
                              anLoopCount[anRollingLoop[0]-1] [anRollingLoop[1]-1]
[anRollingLoop[2]-1] ++;
                              jj++;
                              if (strcmp(szOutputStyle, "-l2") == 0)
                              {
                                 if (bToFile)
                                    fprintf(fpOutput, "%d\t%d\t%d\t|\t%d\t%d\n",
anRollingLoop[0], anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-
apStartPos[0]),jj);
                                 else
                                    printf("%d\t%d\t%d\t|\t%d\t%d\n", anRollingLoop[0],
anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-apStartPos[0]),jj);
                              }
                           }

                        }
                     }
                     else
                     {
                        // otherwise store the start position
                        pStartPos = pPos;
                        apStartPos[3]=pPos;
                     }

                     // move along to the end of the base sequence + the minimum gap
size to look for the next one
                     while (*pPos == cBase)
                     {
                        pPos++;
                        n++;
                     }

                     pLastEndSeq = pPos;

                     // CCCCCCC counts as two strands (CCC)C(CCC)
                     ii += (n+1)/(nBasesInRepeat+nMinGap);

                     if (n >= 2*nBasesInRepeat + nMinGap && n < 3*nBasesInRepeat +
2*nMinGap)
                     {
                        //we have a new loop found
                        //shunt rolling loop count along
                        anRollingLoop[0]=anRollingLoop[1];
                        anRollingLoop[1]=anRollingLoop[2];
                        anRollingLoop[2]=n-2*nBasesInRepeat;

                        //roll array of Startposes
                        apStartPos[0]=apStartPos[1];
```

```
                        apStartPos[1]=apStartPos[2];
                        apStartPos[2]=pPos;
                        apStartPos[3]=pPos+n-nBasesInRepeat;

                        //if we have three loop values, increment the relevant bit of
the array
                        if(anRollingLoop[0] >0)
                        {
                           anLoopCount[anRollingLoop[0]-1] [anRollingLoop[1]-1]
[anRollingLoop[2]-1] ++;
                           jj++;
                           if (strcmp(szOutputStyle, "-l2") == 0)
                           {
                              if (bToFile)
                                 fprintf(fpOutput,"%d\t%d\t%d\t|\t%d\t%d\n",
anRollingLoop[0], anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-
apStartPos[0]),jj);
                              else
                                 printf("%d\t%d\t%d\t|\t%d\t%d\n", anRollingLoop[0],
anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-apStartPos[0]),jj);
                           }
                        }
                     }
                     else if (n>=3*nBasesInRepeat + 2*nMinGap && n <4*nBasesInRepeat +
3*nMinGap)
                     {
                        //we have two new loops found
                        anRollingLoop[0]=anRollingLoop[2];
                        anRollingLoop[1]=(int)(n-3*nBasesInRepeat)/2;
                        anRollingLoop[2]=(n-3*nBasesInRepeat)-anRollingLoop[1];

                        apStartPos[0]=apStartPos[1];
                        apStartPos[1]=pPos;
                        apStartPos[2]=pPos+n-2*nBasesInRepeat-anRollingLoop[2];
                        apStartPos[3]=pPos+n-nBasesInRepeat;

                        //if we have three loop values, increment the relevant bit of
the array
                        if(anRollingLoop[0] >0)
                        {
                           anLoopCount[anRollingLoop[0]-1] [anRollingLoop[1]-1]
[anRollingLoop[2]-1] ++;
                           jj++;
                           if (strcmp(szOutputStyle, "-l2") == 0)
                           {
                              if (bToFile)
                                 fprintf(fpOutput, "%d\t%d\t%d\t|\t%d\t%d\n",
anRollingLoop[0], anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-
apStartPos[0]),jj);
                              else
                                 printf("%d\t%d\t%d\t|\t%d\t%d\n", anRollingLoop[0],
anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-apStartPos[0]),jj);
                     }
                        }
                     }
                     else if (n>4*nBasesInRepeat + 3*nMinGap && n < 5*nBasesInRepeat +
4*nMinGap)
                     {
                        //we have three new loops found. NB this formula is incorrect
for runs equal to ot longer than 5*nBasesInRepeat + 4*nMinGap (normally 19-rare!)
                        anRollingLoop[0]=(int)(n-4*nBasesInRepeat)/3;
                        anRollingLoop[1]=(int)(n-4*nBasesInRepeat)/3;
                        anRollingLoop[2]=(n-4*nBasesInRepeat)-anRollingLoop[1]-
anRollingLoop[0];
                        anLoopCount[anRollingLoop[0]-1] [anRollingLoop[1]-1]
[anRollingLoop[2]-1] ++;

                        apStartPos[0]=pPos;
                        apStartPos[1]=pPos+n-3*nBasesInRepeat-anRollingLoop[1]-
anRollingLoop[2];
                        apStartPos[2]=pPos+n-2*nBasesInRepeat-anRollingLoop[2];
                        apStartPos[3]=pPos+n-nBasesInRepeat;

                        jj++;

                        if (strcmp(szOutputStyle, "-l2") == 0)
                        {
                           if (bToFile)
```

```
                              fprintf(fpOutput, "%d\t%d\t%d\t|\t%d\t%d\n",
anRollingLoop[0], anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-
apStartPos[0]),jj);
                           else
                              printf("%d\t%d\t%d\t|\t%d\t%d\n", anRollingLoop[0],
anRollingLoop[1], anRollingLoop[2],(pPos+nBasesInRepeat-apStartPos[0]),jj);
                        }
//
//                      printf("long limit");
                     }
//                   else if (n>5*nBasesInRepeat + 4*nMinGap)
//                      printf("Exceeded length limit for loops!");

                     pPos += nMinGap;
                  }

                  if (ii >= nRepeatsInSequence)
                  {
                     int nOverlappingSequences = ii + 1 - nRepeatsInSequence;
                     int nSequences = ii/nRepeatsInSequence;

                     nOverlappingSequencesFound += nOverlappingSequences;
                     nSequencesFound += nSequences;
                     nRunsFound ++;

                     cTemp = *pLastEndSeq;
                     *pLastEndSeq = 0;

                     // if we're here, we found a valid sequence of nRepeatsInSequence
or more length - print it out

                     if (strcmp(szOutputStyle, "-c") == 0 || strcmp(szOutputStyle, "-
coord") == 0)
                     {
                        if (bToFile)
                        {
                           if (szPrefix)
                              fprintf(fpOutput, "%s,", szPrefix);

                           fprintf(fpOutput, "%d,%d\n", nChromosomeOffset +(pStartPos
- pStartSequence), nChromosomeOffset + (pLastEndSeq - pStartSequence) - 1);
                        }
                        else
                        {
                           if (szPrefix)
                              printf("%s,", szPrefix);

                           printf("%d,%d\n", nChromosomeOffset + (pStartPos -
pStartSequence), nChromosomeOffset + (pLastEndSeq - pStartSequence) - 1);
                        }
                     }

                     if (strcmp(szOutputStyle, "-d") == 0 || strcmp(szOutputStyle, "-
default") == 0)
                     {
                        if (bToFile)
                           fprintf(fpOutput, "%d-%d\t%d:%d:%d\t%s\n",
nChromosomeOffset + (pStartPos - pStartSequence), nChromosomeOffset + (pLastEndSeq -
pStartSequence) -1 , ii, nOverlappingSequences, nSequences, pStartPos);
                        else
                           printf("%d-%d\t%d:%d:%d\t%s\n", nChromosomeOffset +
(pStartPos - pStartSequence), nChromosomeOffset + (pLastEndSeq - pStartSequence) -1
,ii, nOverlappingSequences, nSequences, pStartPos);
                     }

                     if (strcmp(szOutputStyle, "-p") == 0 || strcmp(szOutputStyle, "-
pos") == 0)
                     {
                        if (bToFile)
                           fprintf(fpOutput, "%d\n", nChromosomeOffset + (pStartPos -
pStartSequence));
                        else
                           printf("%d\n", nChromosomeOffset + (pStartPos -
pStartSequence));
                     }

                     if (strcmp(szOutputStyle, "-g") == 0 ||
                        strcmp(szOutputStyle, "-gene") ==0 ||
                        strcmp(szOutputStyle, "-h") == 0 ||
```

```
                        strcmp(szOutputStyle, "-header") ==0
                        )
                     {
                        if (bToFile)
                           fprintf(fpOutput, "%s\t%s\n", pStartHeader, pStartPos);
                        else
                           printf("%s\t%s\n", pStartHeader, pStartPos);
                     }

                     *pLastEndSeq = cTemp;
                  }

                  pStartPos = NULL;
                  pLastEndSeq = NULL;
               }

               if (strcmp (szOutputStyle, "-number") == 0 || strcmp(szOutputStyle, "-
d") == 0 || strcmp(szOutputStyle, "-default") == 0)
               {
                  if (bToFile)
                     fprintf(fpOutput, "\nFound %d:%d:%d
overlapping:sequences:lines\n", nOverlappingSequencesFound, nSequencesFound,
nRunsFound);
                  else
                     printf("\nFound %d:%d:%d overlapping:sequences:lines\n",
nOverlappingSequencesFound, nSequencesFound, nRunsFound);
               }

               if (strcmp (szOutputStyle, "-loop") == 0 || strcmp(szOutputStyle, "-l")
== 0 )

               {

                  for (ii=0; ii<nMaxGap; ii++)
                  {
                     for (jj=0; jj<nMaxGap; jj++)
                     {
                        for (kk=0; kk<nMaxGap; kk++)
                        {
//                         if(anLoopCount[ii][jj][kk]!=0)
                           {
                              if (bToFile)
                                 fprintf(fpOutput, "%d\t%d\t%d\t%d\n",
ii+1,jj+1,kk+1,anLoopCount[ii][jj][kk]);
                              else
                                 printf("%d\t%d\t%d\t%d\n",
ii+1,jj+1,kk+1,anLoopCount[ii][jj][kk]);
                           }
                        }
                     }
                  }
               }
               if (strcmp (szOutputStyle, "-ll") == 0)
               //abbreviated loop display
               {
//                if (bToFile)
//                   fprintf(fpOutput, "%s\t", pStartHeader);
//                else
//                   printf("%s\t", pStartHeader);

                  for (ii=0; ii<nMaxGap; ii++)
                  {
                     for (jj=0; jj<nMaxGap; jj++)
                     {
                        for (kk=0; kk<nMaxGap; kk++)

                        {
                           if (bToFile)
                              fprintf(fpOutput, "%d\t", anLoopCount[ii][jj][kk]);
                           else
                              printf("%d\t",anLoopCount[ii][jj][kk]);

                        }
                        if (bToFile)
                           fprintf(fpOutput, "\t");
                        else
                           printf("\t");
                     }
                     if (bToFile)
```

```
                        fprintf(fpOutput, "\n");
                     else
                        printf("\n");
                  }
                  if (bToFile)
                     fprintf(fpOutput, "\n");
                  else
                     printf("\n");
               }
            }
         }
      }
   }

   if (bToFile)
      fclose(fpOutput);

   free(szFileContents);
}
```

Appendix C

Selection of quadruplexes conserved in Homo sapiens, Mus Musculus *and* Rattus Norvegicus.

ENSEMBL data for genes known to be homologous between all three species was collected using ENSMART. The 5' untranscribed region (UTR) and 1000 bp upstream was analysed using *quadparser*, and for genes where there were quadruplexes in all three homologues, the results were output with a description of the human gene (take from ENSEMBL), and then a table with the ENSEMBL gene reference numbers and the sequences found.

In most cases shown below, it is apparent that not only is the existence of a quadruplex in this region conserved between all three species, but the sequence is extremely well preserved.

```
++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

OLFACTOMEDIN RELATED ER LOCALIZED PROTEIN ISOFORM 1; NEUROBLASTOMA PROTEIN;
OLFACTOMEDIN   RELATED   ER   LOCALIZED   PROTEIN;   PANCORTIN   1.
[Source:RefSeq;Acc:NM_014279]

ENSG00000130558
                    CCCGCCCAGCCCAGCCCTGCCCAGCCCTGCCC
                    CCCGCCCAGCCCAGCCCTGCCCAGCCCTGCCC
ENSMUSG00000026833
                    GGGCGGCGGCGGGCAGAGGGGGCGCGGGG
ENSRNOG00000009862
                    GGGCGGGCGGCGGCGGGCGGAGGGGGCGCGGGG
                    CCCCTGCCCGCCCTGGTGCCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

MITOGEN-ACTIVATED PROTEIN KINASE 12 (EC 2.7.1.37) (EXTRACELLULAR SIGNAL-
REGULATED KINASE 6) (ERK-6) (ERK5) (STRESS-ACTIVATED PROTEIN KINASE-3)
(MITOGEN-ACTIVATED PROTEIN KINASE P38 GAMMA) (MAP KINASE P38 GAMMA).
[Source:SWISSPROT;Acc:P53778]

ENSG00000185386
                    GGGCGCGGGCGCGGGGCGCGGGGCTGGGCCCGGG
ENSMUSG00000022610
                    GGGCGTGGGGCGCGGGCCGGG
                    GGGCGTGGGGCGCGGGCCGGG
ENSRNOG00000006984
                    GGGCGTGGGGCGCGGGCCGGG

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

```
COUP TRANSCRIPTION FACTOR 2 (COUP-TF2) (COUP-TF II) (APOLIPOPROTEIN AI
REGULATORY PROTEIN-1) (ARP-1). [Source:SWISSPROT;Acc:P24468]

ENSG00000185551
                    CCCTCCCGCGCCCTCTTGCACCC
ENSMUSG00000030551
                    CCCTGAGCCACCCGGGGCGCCCTCCCGCGCCCTCTCGCACCC
ENSRNOG00000010308
                    CCCTGAGCCACCCGGGGCGCCCTCCCGCGCCCTCTTGCACCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

CORTICOSTEROID 11-BETA-DEHYDROGENASE, ISOZYME 2 (EC 1.1.1.146) (11- DH2)
(11-BETA-HYDROXYSTEROID DEHYDROGENASE TYPE 2) (11-BETA-HSD2) (NAD-DEPENDENT
11-BETA-HYDROXYSTEROID DEHYDROGENASE). [Source:SWISSPROT;Acc:P80365]

ENSG00000176387
                    CCCCGCTCCCCGGCCCGGCCCCCGCCCCGCCCCGCCCCAGCCC
ENSMUSG00000031891
                    CCCCAACCAGGCCCCCGCCCCGCCCCAGCCCGGCCCC
ENSRNOG00000017084
                    CCCCAACCCGGCCCCCGCCCCTCCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

PITUITARY HOMEOBOX 1 (HINDLIMB EXPRESSED HOMEOBOX PROTEIN BACKFOOT).
[Source:SWISSPROT;Acc:P78337]

ENSG00000069011
                    CCCAGGCCCACCCCACCCAGCACCCC
                    CCCAGCCCCGGCCCCTGGAGCGCCC
                    CCCAGCCCCGGCCCCTGGAGCGCCC
ENSMUSG00000021506
                    GGGACTGGGGCCCGGGCCGGCCGGG
                    CCCCGAGCCCCGAACCCGGTCCC
                    CCCAGCCCCGGCCCCTGGAGCGCCC
ENSRNOG00000011423
                    GGGACTGGGACCCGGGCCTGCCGGG

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

RETICULON PROTEIN 3 (NEUROENDOCRINE-SPECIFIC PROTEIN-LIKE 2) (NSP-LIKE
PROTEIN II) (NSPLII). [Source:SWISSPROT;Acc:O95197]

ENSG00000133318
                    CCCCTCCCTCTCTCCCGCCCC
ENSMUSG00000024758
                    CCCCTCCCACTCTCCCGCCC
ENSRNOG00000021202
                    CCCCTCCCACTCTCCCGCCCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++

CARTILAGE HOMEOPROTEIN 1 (CART-1). [Source:SWISSPROT;Acc:Q15699]

ENSG00000180318
                    CCCTCCCTCCTCCCACCCACTGGCTCCCTCCCCC
ENSMUSG00000036602
                    GGGTAAGGAGGGAAGATGGGGGTGGGGTGGGGGAGGAGGGGGCCTGGGG
                    CCCTCCCTACTCCCGCCCACTGCTCCCTCCTCCCTCCGCGCCC
ENSRNOG00000004390
                    GGGGGTGGGGTGGGGGAGGAGGGGGCCCTGGGG

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

```
ENSG00000178531
                    GGGGCGGGTGGGGTGGG
ENSMUSG00000048644
                    GGGGCGGGTGGGGTGGGCTGAGCCGGGG
ENSRNOG00000001057
                    GGGGCGGATGGGGTGGGCTGAGCCGGGG

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++


JUNCTION PLAKOGLOBIN (DESMOPLAKIN III). [Source:SWISSPROT;Acc:P14923]

ENSG00000173801
                    CCCACCCCGGCCCGACCCCGACCCGGCCCGGTCAGGCCCC
                    CCCACCCCGGCCCGACCCCGACCCGGCCCGGTCAGGCCCC
ENSMUSG00000001552
                    CCCTCCCGGCCAGACCCGACCCC
ENSRNOG00000015380
                    CCCGGCCAGACCCGACCCCGACCCGGCTCGGCCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++


GABA(A) RECEPTOR-ASSOCIATED PROTEIN; GABA(A)-RECEPTOR-ASSOCIATED PROTEIN.
[Source:RefSeq;Acc:NM_007278]

ENSG00000170296

      CCCCCCGTCCCGGCCCCCCTGGGTTCCCTCAGCCCAGCCCTGTCCAGCCCGGTTCCC
ENSMUSG00000018567
                    CCCCCTGTCCCGGCCCCCCCCCTGGGTTCCCTCAGCCC
ENSRNOG00000017417
                    CCCGGCCCCCCCCCTGGGTTCCCTCAGCCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++


GUANINE NUCLEOTIDE-BINDING PROTEIN G(I), ALPHA-2 SUBUNIT (ADENYLATE CYCLASE-
INHIBITING G ALPHA PROTEIN). [Source:SWISSPROT;Acc:P04899]

ENSG00000114353
                    GGGTCGGGCGGGGCCGAGCCGGGCCGTGGGCCGTGTGGGGGCCGGG
ENSMUSG00000032562
                    GGGTCGGGCGGGGCCGAGCCGGGCCGTGGGCCGTGTGGGGGCCAGGCCGGG
ENSRNOG00000016592
                    GGGTCGGGCGGGGCCGAGCCGGGCCGTGGGCCGTGTGGGGGCCAGGCCGGG

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++


RAS-RELATED PROTEIN RAB-13. [Source:SWISSPROT;Acc:P51153]

ENSG00000143545
                    CCCCACCCCTCCCCCGGCTCCCCC
ENSMUSG00000027935
                    CCCGACCCCTCCCCCGGCGCCCCC
ENSRNOG00000016733
                    CCCGACCCCTCCCCCGGCGCCCCC

++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++++
```

REFERENCES

1. Watson, J.D. and Crick, F.H.C., 'A structure for deoxyribose nucleic acid'. *Nature*, **1953**. 171: pp. 737-8.
2. Zamenhof, S., Brawerman, G., and Chargaff, E., 'On the desoxypentose nucleic acid from several microorganisms'. *Biochim. et Biophys. Acta*, **1952**. 9: pp. 402.
3. Franklin, R.E. and Gosling, R.G., 'Molecular configuration in sodium Thymonucleate'. *Nature*, **1953**. 171: pp. 740-1.
4. Wilkins, M.H.F., Stokes, A.R., and Wilson, H.R., 'Molecular structure of deoxypentose nucleic acids'. *Nature*, **1953**. 171: pp. 738-740.
5. Lodish, H., Baltimore, D., Berk, A., Zipursky, S.L., Matsudaira, P., and Darnell, J., *Molecular Cell Biology*. Third ed. 1996: Scientific American Books.
6. Sayle, R.A. and Milner-White, E.J., 'RasMol: Biomolecular graphics for all'. *Trends Bioc. Sci.*, **1995**. 20: pp. 374-6.
7. Bang, I., 'Untersuchungen über die Guanylsäre'. *Biochemisce Zeitschrift*, **1910**. 26: pp. 293-311.
8. Gellert, M., Lipsett, M.N., and Davies, D.R., 'Helix formation by guanylic acid'. *Proc. Natl. Acad. Sci. USA*, **1962**. 48: pp. 2013-8.
9. Guschlbauer, W., Chantot, J.F., and D., T., 'Four-stranded nucleic structures 25 years later: from guanosine gels to telomere DNA'. *J. Biomol. Struct. Dyn.*, **1990**. 8: pp. 491-511.
10. Sasisekharan, V., Zimmerman, S.B., and Davies, D.R., 'The structure of helical 5'-guanosine monophosphate'. *J. Mol. Biol.*, **1975**. 92: pp. 171-9.
11. Sen, D. and Gilbert, W., 'Guanine quartet structures'. *Meth. Enzymol.*, **1992**. 211: pp. 191-199.
12. Simonsson, T., 'G-Quadruplex DNA structures - variations on a theme'. *Biol. Chem.*, **2001**. 382: pp. 621-628.
13. Zimmerman, S.B., 'An acid structure for polyriboguanylic acid observed by X-ray diffraction'. *Biopolymers*, **1975**. 14: pp. 889-890.
14. Zimmerman, S.B., Cohen, G.H., and Davies, D.R., 'X-ray fiber diffraction and model-building study of polyguanylic acid and polyinosinic acid'. *J. Mol. Biol.*, **1975**. 92: pp. 181-192.
15. Arnott, S., Chandrasekaran, R., and Marttila, C.M., 'Structures for polyinosinic acid and polyguanylic acid'. *Biochem. J.*, **1974**. 141: pp. 537-543.
16. Laughlan, G., Murchie, A.I., Norman, D.G., Moore, M.H., Moody, P.C., Lilley, D.M., and Luisi, B., 'The high-resolution crystal structure of a parallel-stranded quanine tatraplex'. *Science*, **1994**. 265: pp. 520-524.
17. Williamson, J.R., 'G-quartet structures in telomeric DNA'. *Ann. Rev. Biophys. Biomol. Struct.*, **1994**. 23: pp. 703-30.
18. Deng, J., Xiong, Y., and Sundaralingam, M., 'X-ray analysis of an RNA tetraplex $(UGGGGU)_4$ with divalent Sr^{2+} ions at subatomic resolution (0.61Å)'. *Proc. Natl. Acad. Sci. USA*, **2001**. 98(24): pp. 13665-13670.
19. Datta, B., Schmitt, C., and Armitage, B.A., 'Formation of a PNA_2-DNA_2 hybrid quadruplex'. *J. Am. Chem. Soc.*, **2002**.
20. Krishnan-Ghosh, Y., Stephens, E., and Balasubramanian, S., 'A PNA_4 quadruplex'. *J. Am. Chem. Soc.*, **2004**. 126: pp. 5944-5.

21. Randazzo, A., Esposito, V., Ohlenschläger, O., Ramachandran, R., and Mayol, L., 'NMR solution structure of a parallel LNA quadruplex'. *Nucl. Ac. Res.*, **2004**. 32(10): pp. 3083-3092.
22. Phillips, K., Dauter, Z., Murchie, A.I., Lilley, D.M., and Luisi, B., 'The crystal structure of a parallel-stranded guanine tetraplex at 0.95Å resolution'. *J. Mol. Biol.*, **1997**. 273: pp. 171-182.
23. Marathias, V.M. and Bolton, P.H., 'Determinants of DNA quadruplex structural type: Sequence and potassium binding'. *Biochemistry*, **1999**. 38(14): pp. 4355-4364.
24. Sen, D. and Gilbert, W., 'A sodium-potassium switch in the formation of four-stranded G4-DNA'. *Nature*, **1990**. 344: pp. 410-4.
25. Phan, A.T. and Patel, D.J., 'Two-repeat human telomeric d(TAGGGTTAGGGT) Sequence Forms Interconverting Parallel and Antiparallel G-Quadruplexes in Solution: Distinct Topologies, Thermodynamic Properties, and Folding/Unfolding Kinetics'. *J. Am. Chem. Soc.*, **2003**. 125(49): pp. 15021-7.
26. Wang, Y. and Patel, D.J., 'Solution structure of the *Oxytricha* telomeric repeat d[$G_4(T_4G_4)_3$] G tetraplex'. *J. Mol. Biol.*, **1995**. 251(1): pp. 76-94.
27. Wang, Y. and Patel, D.J., 'Solution structure of the human telomeric repeat d[$AG_3(T_2AG_3)_3$] G-tetraplex'. *Structure*, **1993**. 1: pp. 263-282.
28. Macaya, R.F., Schultze, P., Smith, F.W., Roe, J.A., and Feigon, J., 'Thrombin-binding DNA aptamer forms a unimolecular quadruplex structure in solution'. *Proc. Natl. Acad. Sci. USA*, **1993**. 90: pp. 3745-9.
29. Parkinson, G.H., Lee, M.P.H., and Neidle, S., 'Crystal structure of parallel quadruplexes from human telomeric DNA'. *Nature*, **2002**. 417: pp. 876-880.
30. Sundquist, W.I. and Klug, A., 'Telomeric DNA dimerizes by formation of guanine tetrads between hairpin loops'. *Nature*, **1989**. 342: pp. 825-9.
31. Sen, D. and Gilbert, W., 'Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis'. *Nature*, **1988**. 334: pp. 364-6.
32. Laughlan, G., Murchie, A.I.H., Norman, D.G., Moore, M.H., Moody, P.C., Lilley, D.M., and Luisi, B., 'The high-resolution crystal structure of a parallel-stranded guanine tetraplex'. *Science*, **1994**. 265: pp. 520-524.
33. Jin, R., Gaffney, B.L., Wang, C., Jones, R.A., and Breslauer, K.J., 'Thermodynamics and structure of a DNA tetraplex: a spectroscopic and calorimetric study of the tetramolecular complexes of d(TG_3T) and d($TG_3T_2G_3T$)'. *Proc. Natl. Acad. Sci. USA*, **1992**. 89: pp. 8832-6.
34. Shiber, M.C., Braswell, E.H., Klump, H., and Fresco, J.R., 'Duplex-tetraplex equilibrium between a hairpin and two interacting hairpins of d$(A\text{-}G)_{10}$ at neutral pH'. *Nucl. Ac. Res.*, **1996**. 24: pp. 5004-5012.
35. Hardin, C.C., Henderson, E., Watson, T., and Prosser, J.K., 'Monovalent cation induced structural transitions in telomeric DNA: G-DNA folding intermediates'. *Biochemistry*, **1991**. 30: pp. 4460-4472.
36. Wang, Y. and Patel, D.J., 'Solution structure of the *Tetrahymena* telomeric repeat d[(T_2G_4) G-tetraplex'. *Structure*, **1994**. 2(12): pp. 1141-56.
37. Smith, F.W. and Feigon, J., 'Quadruplex structure of *Oxytrichia* telomeric DNA oligonucleotides'. *Nature*, **1992**. 356(6365): pp. 164-8.

38. Keniry, M.A., 'Quadruplex structures in nucleic acids'. *Biopolymers*, **2000**. 56(3): pp. 123-146.
39. Blackburn, E.H., 'Telomeres and their synthesis'. *Science*, **1990**. 249: pp. 489-90.
40. Crnugelj, M., Sket, P., and Plavec, J., 'Small change in G-rich sequence, a dramatic change in topology: new dimeric G-quadruplex folding motif with unique loop orientations'. *J. Am. Chem. Soc.*, **2003**. 125(26): pp. 7866-7871.
41. Bonnal, S., Schaeffer, C., Créancier, L., Clamens, S., Moine, H., Prats, A.-C., and Vagner, S., 'A Single Internal Ribosome Entry Site Containing a G Quartet RNA Structure Drives Fibroblast Growth Factor 2 Gene Expression at Four Alternative Translation Initiation Codons'. *J. Biol. Chem.*, **2003**. 278(41): pp. 39330-6.
42. Risitano, A. and Fox, K.R., 'The stability of intramolecular DNA quadruplexes with extended loops forming inter- and intra-loop duplexes'. *Org. Biomol. Chem.*, **2003**. 1: pp. 1852-5.
43. Cech, T.R., 'Life at the end of the chromosome: telomeres and telomerase'. *Angew. Chem. Int. Ed. (Engl.)*, **2000**. 39: pp. 34-43.
44. Blackburn, E., 'Structure and function of telomeres'. *Nature*, **1991**. 350: pp. 569-573.
45. Horvath, M.P., Schweiker, V.L., Bevilacqua, J.M., Ruggles, J.A., and Schultz, S.C., 'Crystal Structure of the *Oxytricha nova* telomere end binding protein complexed with single strand DNA'. *Cell*, **1998**. 95: pp. 963-974.
46. Griffith, J.D., Comeau, L., Rosenfield, S., Stansel, R.M., Bianchi, A., Moss, H., and de Lange, T., 'Mammalian telomeres end in a large duplex loop'. *Cell*, **1999**. 97: pp. 503-514.
47. Neidle, S. and Parkinson, G.H., 'The structure of telomeric DNA'. *Curr. Op. Struct. Biol.*, **2003**. 13(3): pp. 275-263.
48. Greider, C. and Blackburn, E., 'Identification of a specific telomere terminal transferase activity in Tetrahymena extracts'. *Cell*, **1985**. 43: pp. 405-413.
49. O'Reilly, M., Teichmann, S., and Rhodes, D., 'Telomerases'. *Curr. Op. Struct. Biol.*, **1999**. 9: pp. 56-65.
50. Wright, W.E., 'Normal human chromosomes have long G-rich telomeric overhangs at one end'. *Genes Dev.*, **1997**. 11: pp. 2801-9.
51. Blackburn, E.H., 'Structure and function of telomeres'. *Nature*, **1991**. 350: pp. 569-573.
52. Schaffitzel, C., Berer, I., Postberg, J., Hanes, J., Lipps, H.J., and Plückthun, A., '*In vitro* generated antibodies specific for telomeric guanine-quadruplex DNA react with *Stylonychia lemnae* macronuclei'. *Proc. Natl. Acad. Sci. USA*, **2001**. 98(15): pp. 8572-8577.
53. Gehring, K., Leroy, J., and Guéron, M., 'A tetrameric DNA structure with protonated cytosine•cytosine base pairs'. *Nature*, **1993**. 363: pp. 499-510.
54. Risitano, A. and Fox, K.R., 'Stability of intramolecular DNA quadruplexes: comparison with DNA duplexes'. *Biochemistry*, **2003**. 42(21): pp. 6507-13.

55. Li, W., Miyoshi, D., Nakano, S., and Sugimoto, N., 'Structural competition involving G-quadruplex DNA and its complement'. *Biochemistry*, **2003**.
56. Li, W., Wu, P., Ohmichi, T., and Sugimoto, N., 'Characterization and thermodynamic properties of quadruplex/duplex competition'. *FEBS Letters*, **2002**. 526: pp. 77-81.
57. Phan, A.T. and Mergny, J.-L., 'Human telomeric DNA: G-quadruplex, i-motif and Watson–Crick double helix'. *Nucl. Ac. Res.*, **2002**. 30(21): pp. 4618-25.
58. Miyoshi, D., Nakao, A., and Sugimoto, N., 'Molecular crowding regulates the structural switch of the DNA G-quadruplex'. *Biochemistry*, **2002**.
59. Rangan, A., Fedoroff, O.Y., and Hurley, L.H., 'Induction of duplex to G-quadruplex transition in the c-myc promoter region by a small molecule'. *J. Biol. Chem.*, **2001**. 276(7): pp. 4640-4646.
60. Mergny, J.-L., Riou, J.-F., Mailliet, P., Teulade-Fichou, M.-P., and Gilson, E., 'Natural and pharmacological regulation of telomerase'. *Nucl. Ac. Res.*, **2002**. 30(4): pp. 839-865.
61. Fletcher, T.M., Sun, D., Salazar, M., and Hurley, L.H., 'Effect of DNA secondary structure on human telomerase activity'. *Biochemistry*, **1998**. 37(16): pp. 5536-5541.
62. Mergny, J.-L., Lacroix, L., Teulade-Fichou, M.-P., Hounsou, C., Guittat, L., Hoarau, M., Arimondo, P.B., Vigneron, J., Lehn, J., Riou, J.-F., Garestier, T., and Hélène, C., 'Telomerase inhibitors based on quadruplex ligands selected by a fluorescence assay'. *Proc. Natl. Acad. Sci. USA*, **2001**. 98(6): pp. 3062-7.
63. Kang, C., Zhang, X., Ratliff, R., Moyzis, R., and Rich, A., 'Crystal structure of four-stranded *Oxytricha* telomeric DNA'. *Nature*, **1992**. 356: pp. 126-131.
64. Balagurumoorthy, P., Brahmachari, S.K., Mohanty, D., Bansal, M., and Sasiskharan, V., 'Hairpin and parallel quartet structure for telomeric sequences'. *Nucl. Ac. Res.*, **1992**. 20: pp. 4061-7.
65. Mergny, J.-L. and Maurizot, J., 'Fluorescence Resonance Energy Transfer as a probe for G-quartet formation by a telomeric repeat'. *ChemBioChem*, **2001**. 2: pp. 124-132.
66. Ying, L.M., Green, J.J., Li, H.T., Klenerman, D., and Balasubramanian, S., 'Studies on the structure and dynamics of the human telomeric G quadruplex by single-molecule fluorescence resonance energy transfer'. *Proc. Natl. Acad. Sci. USA*, **2003**. 100(25): pp. 14629-34.
67. Kerwin, S.M., 'G-quadruplex DNA as a target for drug design'. *Curr. Pharm. Des.*, **2000**. 6: pp. 441-71.
68. Neidle, S. and Parkinson, G.H., 'Telomere maintenance as a target for anticancer drug discovery'. *Nat. Rev. Drug Disc.*, **2002**. 1: pp. 383-393.
69. Kim, N.W., Piatszek, M.A., Prowse, K.R., Harley, C.B., West, M.D., Ho, P.L.C., Coviello, G.M., Wright, W.E., Weinrich, S.L., and Shay, J.W., 'Specific Association of Human Telomerase Activity with Immortal Cells and Cancer'. *Science*, **1994**. 266: pp. 2011-2015.
70. Han, H., Hurley, L.H., and Salazar, M., 'A DNA polymerase stop assay for G-quadruplex-interactive compounds'. *Nucl. Ac. Res.*, **1999**. 27(2): pp. 537-542.

71. Anantha, N.V., Azam, M., and Sheardy, R.D., 'Porphyrin binding to quadruplexed T_4G_4'. *Biochemistry*, **1998**. 37(9): pp. 2709-2714.
72. Haq, I., Trent, J.O., Chowdhry, B.Z., and Jenkins, T.C., 'Intercalative G-tetraplex stabilization of telomeric DNA by a cationic prophyrin'. *J. Am. Chem. Soc.*, **1999**. 121(9): pp. 1768-1779.
73. Koeppel, F., Riou, J.-F., Laoui, A., Mailliet, P., Arimondo, P.B., Labit, D., Petitgenet, O., Hélène, C., and Mergny, J.-L., 'Ethidium derivatives bind to G-quartets, inhibit telomerase and act as fluorescent probed for quadruplexes'. *Nucl. Ac. Res.*, **2001**. 29(5): pp. 1087-1096.
74. Fedoroff, O.Y., Salazar, M., Han, H., Chemeris, S.M., Kerwin, S.M., and Hurley, L.H., 'NMR-Based model of a telomerase-inhibiting compound bound to G-quadruplex DNA'. *Biochemistry*, **1998**. 37(36): pp. 12367-12374.
75. Han, H., Cliff, C.L., and Hurley, L.H., 'Accelerated assembly of G-quadruplex structures by a small molecule'. *Biochemistry*, **1999**. 38: pp. 6981-6.
76. Han, H., Bennett, R.J., and Hurley, L.H., 'Inhibition of unwinding of G-Quadruplex structures by Sgs1 helicase in the presence of N,N'-bis[2-(1-piperidino)ethyl]-3,4,9,10-perylenetetracarboxylic diimide, a G-quadruplex-interactive ligand'. *Biochemistry*, **2000**. 39: pp. 9311-9316.
77. Chen, Q., Kuntz, I.D., and Shafer, R.H., 'Spectroscopic recognition of guanine dimeric hairpin quadruplexes by a carbocyanine dye'. *Proc. Natl. Acad. Sci. USA*, **1996**. 93: pp. 2635-9.
78. Read, M., Harrison, R.J., Romagnoli, B., Tanious, F.A., Gowan, S.H., Reszka, A.P., Wilson, W.D., Kelland, L.R., and Neidle, S., 'Structure-based design of selective and potent Gquadruplex-mediated telomerase inhibitors'. *Proc. Natl. Acad. Sci. USA*, **2001**. 98(9): pp. 4844-9.
79. Read, M.A. and Neidle, S., 'Structural characterization of a guanine-quadruplex ligand complex'. *Biochemistry*, **2000**. 39: pp. 13422-13432.
80. Kim, M.-Y., Vankayalapati, H., Shin-ya, K., Wierzba, K., and Hurley, L.H., 'Telomestatin, a potent telomerase inhibitor that interacts quite specifically with the human telomeric intramolecular G-quadruplex'. *J. Am. Chem. Soc.*, **2002**.
81. Schouten, J.A., Ladame, S., Mason, S.J., Cooper, M.A., and Balasubramanian, S., 'G-Quadruplex-specific peptide-hemiocyanine ligands by partial combinatorial selection'. *J. Am. Chem. Soc.*, **2003**. 125(19): pp. 5594-5.
82. Yin, F., Liu, J., and Peng, X., 'Triethylene tetraamine: a novel telomerase inhibitor'. *Bioorg. Med. Chem. Lett.*, **2003**. 13: pp. 3923-6.
83. Sun, H., J.K., K., Hickson, I.D., and Maizels, N., 'The Bloom's syndrome helicase unwinds G4 DNA'. *J. Biol. Chem.*, **1998**. 273: pp. 27587-92.
84. Fry, M. and Leob, L.A., 'Human Werner syndrome DNA helicase unwinds tetrahelical structures of the fragile X syndrome repeat sequence $d(CGG)_n$'. *J. Biol. Chem.*, **1999**. 274: pp. 12797-802.
85. Li, J.-L., Harrison, R.J., Reszka, A.P., Brosh, R.M., Jr., Bohr, V.A., Neidle, S., and Hickson, I.D., 'Inhibition of the Bloom's and Werner's Syndrome Helicases by G-Quadruplex Interacting Ligands'. *Biochemistry*, **2001**.

86. Giraldo, R. and Rhodes, D., 'The yeast telomere-binding protein RP1 binds to and promotes the formation of DNA quadruplexes in telomeric DNA'. *The EMBO Journal*, **1994**. 13(10): pp. 2411-2420.
87. Giraldo, R., Suzuki, M., Chapman, L., and Rhodes, D., 'Promotion of parallel DNA quadruplexes by a yeast telomere binding protein: A circular dichroism study'. *Proc. Natl. Acad. Sci. USA*, **1994**. 91: pp. 7658-7562.
88. Laporte, L. and Thomas, G.J., Jr., 'Structural basis of DNA recognition and mechanism of quadruplex formation by the β subunit of the *Oxytricha* telomere-binding protein'. *Biochemistry*, **1998**. 37: pp. 1327-1335.
89. Sarig, G., Weisman-Shomer, P., Erlitzki, R., and Fry, M., 'Purification and characterization of qTBP42, a new single-stranded and quadruplex telomeric DNA-binding protein from rat hepatocytes'. *J. Biol. Chem.*, **1997**. 272(7): pp. 4474-4482.
90. Ying, J., Bradley, R.K., Jones, L.B., Reddy, M.S., Colbert, D.T., Smalley, R.E., and Hardin, S.H., 'Guanine-rich telomeric sequences stimulate DNA polymerase activity in vitro'. *Biochemistry*, **1999**. 38: pp. 16461-8.
91. Zahler, A.M., Williamson, J.R., Cech, T.R., and Prescott, D.M., 'Inhibition of telomerase by G-quartet DNA structures'. *Nature*, **1991**. 350: pp. 718-720.
92. Isalan, M., Patel, S.D., Balasubramanian, S., and Choo, Y., 'Selection of zinc fingers that bind single-stranded telomeric DNA in the G-quadruplex confirmation'. *Biochemistry*, **2001**. 40: pp. 830-836.
93. Patel, S., 'Studies on a designed G-quadruplex binding protein that inhibits human telomerase'. *PhD Thesis*, **2000**. Univ. of Cambridge.
94. Pavletich, N.P. and Pabo, C.O., 'Zinc finger - DNA recognition: crystal structure of a Zif268-DNA complex at 2.1 Å'. *Science*, **1991**. 252: pp. 809-817.
95. Bock, L.C., Griffin, L.C., Latham, J.A., Vermaas, E.H., and Toole, J.J., 'Selection of single-stranded DNA molecules that bind and inhibit human thrombin'. *Nature*, **1992**. 355: pp. 564-566.
96. Wang, K.Y., McCurdy, S., Shea, R.G., Swminathan, S., and Bolton, P.H., 'A DNA Aptamer which binds to and inhibits thrombin exhibits a new structural motif for DNA'. *Biochemistry*, **1993**. 32(8): pp. 1899-1904.
97. Wyatt, J.R., Vickers, T.A., Robertson, J.R., Buckheit, R.W., Jr., Klimkait, T., De-Barts, E., Davis, P.W., Rayner, B., Imbach, J.L., and Ecker, D.J., 'Combinatorially selected guanosine-quartet structure is a potent inhibitor of human immunodefficiency virus envelope-mediated cell fusion'. *Proc. Natl. Acad. Sci. USA*, **1994**. 91(4): pp. 1356-1360.
98. Matsugami, A., Okuizumi, T., Uesugi, S., and Katahira, M., 'Intramolecular higher-order packing of parallel quadruplexes comprising a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G heptad of GGA triplet repeat DNA'. *J. Biol. Chem.*, **2003**. 278(30): pp. 28147-28153.
99. Matsugami, A., Ouhashi, K., Kanagawa, M., Liu, H., Kanagawa, S., Uesugi, S., and Katahira, M., 'An intramolecular quadruplex of $(GGA)_4$ triplet repeat DNA with a G:G:G:G tetrad and a G(:A):G(:A):G(:A):G

heptad, and its dimeric interaction'. *J. Mol. Biol.*, **2001**. 313(2): pp. 255-269.
100. Smirnov, I. and Shafer, R.H., 'Effect of loop sequence and size on DNA aptamer stability'. *Biochemistry*, **2000**. 39: pp. 1462-8.
101. Patel, P.K. and Hosur, R.V., 'NMR observation of T-tetrads in a parallel stranded DNA quadruplex formed by Saccharomyces cerevisiae telomere repeats'. *Nucl. Ac. Res.*, **1999**. 27(12): pp. 2457-2464.
102. Cáceres, C., Wright, G., Gouyette, C., Parkinson, G.H., and Subirana, J.A., 'A thymine tetrad in d(TGGGGT) quadruplexes stabilized with Tl^+/Na^+ Ions'. *Nucl. Ac. Res.*, **2004**. 32(3): pp. 1097-1102.
103. Patel, P.K., Koti, A.S.R., and Hosur, R.V., 'NMR studies on truncated sequences of human telomeric DNA: observation of a novel A-tatrad.' *Nucl. Ac. Res.*, **1999**. 27(19): pp. 3836-3843.
104. Krishnan-Ghosh, Y., Liu, D., and Balasubramanian, S., 'Formation of an interlocked quadruplex dimer by d(GGGT)'. *J. Am. Chem. Soc.*, **2004**. 125: pp. 11009-11016.
105. Costa, L.T., Kerkmann, M., Hartmann, G., Endres, S., Bisch, P.M., Heckl, W.M., and Thalhammer, S., 'Structural studies of oligonucleotides containing G-quadruplex motifs using AFM'. *Biochem. Biophys. Res. Comm.*, **2004**. 313(4): pp. 1065-72.
106. Liu, D. and Balasubramanian, S., 'A proton fuelled DNA nanomachine'. *Angew. Chem. Int. Ed. (Engl.)*, **2003**. 42: pp. 5734-6.
107. Fry, M. and Leob, L.A., 'The Fragile X syndrome $d(CGG)_n$ nucleotide repeats form a stable tetrahelical structure'. *Proc. Natl. Acad. Sci. USA*, **1994**. 91: pp. 4950-4.
108. Fojtik, P., Kejnovska, I., and Vorlickova, M., 'The guanine-rich fragile X chromosome repeats are reluctant to form tetraplexes'. *Nucl. Ac. Res.*, **2004**. 32(1): pp. 298-306.
109. Saha, T. and Usdin, K., 'Tetraplex formation by the progressive myoclonus epilepsy type-1 repeat: implications of instability in the repeat expansion diseases'. *FEBS Letters*, **2001**. 491: pp. 184-187.
110. Castati, P., Chen, X., Moyzis, R.K., Bradbury, E.M., and Gupta, G., 'Structure-function correlations of the insulin-linked polymorphic region'. *J. Mol. Biol.*, **1996**. 264(3): pp. 534-545.
111. Weitzmann, M.N., Woodford, K.J., and Usdin, K., 'The mouse Ms6-hm hypervariable microsatellite forms a hairpin and two unusual tetraplexes'. *J. Biol. Chem.*, **2002**. 273(46): pp. 30742-9.
112. Facchini, L.M. and Penn, L.Z., 'The molecular role of Myc in growth and transformation: recent discoveries lead to new insights'. *Faseb Journal*, **1998**. 12: pp. 633-651.
113. Simonsson, T., Pecinka, P., and Kubista, M., 'DNA tetraplex formation in the control region of *c-myc*'. *Nucl. Ac. Res.*, **1998**. 26: pp. 1167-72.
114. Sun, D., Pourpak, A., Beetz, K., and Hurley, L.H., 'Direct evidence for the formation of G-quadruplex in the proximal promoter region of the RET protooncogene and its targetting with a small molecule to repress RET protooncogene transcription'. *Clin. Cancer Res. (Supplement)*, **2003**. 9(16): pp. A218.
115. Cogoi, S., Quadrifoglio, F., and Xodo, L.E., 'G-rich oligonucleotide inhibits the binding of a nuclear protein to the Ki-*ras* promoter and

strongly reduces cell growth in human carcinoma pancreatic cells'. *Biochemistry*, **2004**. 43: pp. 2512-2523.
116. Christansen, J., Kofod, M., and Nielsen, F.C., 'A guanosine quadruplex and two stable hairpins flank a major cleavage site in insulin-like growth factor II mRNA'. *Nucl. Ac. Res.*, **1994**. 22(25): pp. 5709-16.
117. Slamon, D.J., deKernion, J.B., Verma, I.M., and Cline, M.J., 'Expression of cellular oncogenes in human malignancies'. *Science*, **1984**. 224(4646): pp. 256-62.
118. Marcu, K.B., Bossone, S.A., and Patel, A.J., 'myc function and regulation'. *Ann. Rev. Biochem.*, **1992**. 61: pp. 809-860.
119. Davis, T.L., Firulli, A.B., and Kinniburgh, A.J., 'Ribonucleoprotein and protein factors bind to an H-DNA-forming c-myc DNA element: possible regulators of the c-myc gene'. *Proc. Natl. Acad. Sci. USA*, **1989**. 86(24): pp. 9682-6.
120. Boles, T.C. and Hogan, M.E., 'DNA structure equilibria in the Human c-myc gene'. *Biochemistry*, **1987**. 26: pp. 367-376.
121. Siebenlist, U., Hennighausen, L., Battey, J., and Leder, P., 'Chromatin structure and protein binding in the putative regulatory region of the c-myc gene in Burkitt lymphoma'. *Cell*, **1984**. 37(2): pp. 381-391.
122. Cooney, M., Czernuszewicz, G., Postel, E.H., Flint, S.J., and Hogan, M.E., 'Site-specific oligonucleotide binding represses transcription of the human c-myc gene in vitro'. *Science*, **1988**. 241(4864): pp. 456-9.
123. Postel, E.H., Flint, S.J., Kessler, D.J., and Hogan, M.E., 'Evidence that a triplex-forming oligodeoxyribonucleotide binds to the c-myc promoter in HeLa cells, thereby reducing c-myc mRNA levels.' *Proc. Natl. Acad. Sci. USA*, **1991**. 88(18): pp. 8227-31.
124. Olivas, W.M. and Maher, L.J., 3rd, 'Competitive triplex/quadruplex equilibria involving guanine-rich oligonucleotides.' *Biochemistry*, **1995**. 34(1): pp. 278-84.
125. Siddiqui-Jain, A., Grand, C.L., Bearss, D.J., and Hurley, L.H., 'Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress *c-MYC* transcription'. *Proc. Natl. Acad. Sci. USA*, **2002**. 99(18): pp. 11593-8.
126. Grand, C.L., Han, H., Muñoz, R.M., Weitman, S., Von Hoff, D.D., Hurley, L.H., and Bearss, D.J., 'The cationic porphyrin TMPyP4 down-regulates *c-MYC* and human telomerase reverse transcriptase expression and inhibits tumor growth *in vivo*'. *Molecular Cancer Therapeutics*, **2002**. 1: pp. 565-573.
127. Seenisamy, J., Rezler, E.M., Powell, T.J., Tye, D., Gokhale, V., Joshi, C.S., Siddiqui-Jain, A., and Hurley, L.H., 'The dynamic character of the G-quadruplex element in the c-MYC promoter and modification by TMPyP4'. *J. Am. Chem. Soc.*, **2004**. 126(28): pp. 8702-9.
128. Phan, A.T., Modi, Y.S., and Patel, D.J., 'Propellor-type parallel-stranded G-quadruplexes in the human *c-myc* promoter'. *J. Am. Chem. Soc.*, **2004**.
129. Mergny, J.-L., Phan, A., and Lacroix, L., 'Following G-quartet formation by UV-spectroscopy'. *FEBS Letters*, **1998**. 435: pp. 74-8.
130. Rost, B. and Sander, C., 'Prediction of protein secondary structure at better than 70% accuracy'. *J. Mol. Biol.*, **1993**. 232: pp. 584-599.

131. Petraccone, L., Erra, E., Esposito, V., Randazzo, A., Mayol, L., Nasti, L., Barone, G., and Giancola, C., 'Stability and structure of DNA sequences forming quadruplexes containing four G-tetrads with different topological arrangements'. *Biochemistry*, **2004**. 43: pp. 4877-4884.
132. Fojtik, P. and Vorlickova, M., 'The fragile X chromosome (GCC) repeat folds into a DNA tetraplex at neutral pH'. *Nucl. Ac. Res.*, **2001**. 29(22): pp. 4684-4690.
133. da Silva, M.W., 'Association of DNA quadruplexes through G:C:G:C tetrads. Solution structure of d(GCGGTGGAT)'. *Biochemistry*, **2003**. 42: pp. 14356-14365.
134. Escaja, N., Gelpí, J.L., Orozco, M., Rico, M., Pedroso, E., and González, C., 'Four-stranded DNA structure stabilized by a novel G:C:A:T tetrad'. *J. Am. Chem. Soc.*, **2003**. 125(19): pp. 5654-5662.
135. Risitano, A. and Fox, K.R., 'Influence of loop size on the stability of intramolecular G-quadruplexes'. *Nucl. Ac. Res.*, **2004**. 32(8): pp. 2598-2606.
136. Maxam, A.M. and Gilbert, W., 'A new method for sequencing DNA'. *Proc. Natl. Acad. Sci. USA*, **1977**. 74: pp. 560-4.
137. Maxam, A.M. and Gilbert, W., 'Sequencing end-labelled DNA with base-specific chemical cleavages'. *Meth. Enz.*, **1980**. 65: pp. 499-510.
138. Li, T.H., Liu, D., Chen, J., Lee, A.H.F., Qi, J., and Chan, A.S.C., 'Construction of Circular Oligodeoxyribonucleotides on the New Structural Basis of i-Motif'. *J. Am. Chem. Soc.*, **2001**.
139. Berova, N., Nakanishi, K., and Woody, R.W., *Circular dichroism: principles and applications*. 2nd ed. 2000, New York; Chichester: Wiley-VCH.
140. Lobley, A., Whitmore, L., and Wallace, B.A., 'Dichroweb - on line circular dichroism analysis'. *www.cryst.bbk.ac.uk/cdweb/home.html*, **2002**.
141. Kypr, J., Fialova, M., Chladkova, J., Tumova, M., and Vorlickova, M., 'Conserved guanine-guanine stacking in tetraplex and duplex DNA'. *Eur. Biophys. J.*, **2001**. 30: pp. 555-558.
142. Kypr, J. and Vorlickova, M., 'Circular dichroism spectroscopy reveals invariant conformation of guanine runs in DNA'. *Biopolymers*, **2002**. 67: pp. 275-277.
143. Williamson, J.R., Raghuraman, M.K., and Cech, T.R., 'Monovalent cation-induced structure of telomeric DNA: The G-quartet model'. *Cell*, **1989**. 59: pp. 871-880.
144. Hazel, P., Huppert, J.H., Balasubramanian, S., and Neidle, S., 'Loop-length dependent folding of G-quadruplexes'. *J. Am. Chem. Soc.*, **November 23, 2004**: pp. 10.1021/ja045154j.
145. Spackova, N., Berger, I., and Sponer, J., 'Nanosecond molecular dynamics simulatitions of parallel and antiparallel quadruplex DNA molecules'. *J. Am. Chem. Soc.*, **1999**. 121(23): pp. 5519-34.
146. Stefl, R., Cheatham, T.E., III, Spackova, N., Fardna, E., Berger, I., Koca, J., and Sponer, J., 'Formation pathways of a guanine-quadruplex DNA revealed by molecular dynamics and thermodynamic analysis of the substrates'. *Biophysical Journal*, **2003**. 85: pp. 1787-1804.

147. Malliavin, T.E., Gau, J., Snoussi, K., and Leroy, J.-L., 'Stability of the *i*-motif is related to the interactions of phosphodiester backbones'. *Biophysical Journal*, **2003**. 84: pp. 3838-3847.
148. Stefl, R., Spackova, N., Berger, I., Koca, J., and Sponer, J., 'Molecular dynamics of DNA quadruplex molecules containing inosine, 6-thioguanine and 6-thiopurine'. *Biophysical Journal*, **2001**. 80(1): pp. 455-468.
149. Chowdury, S. and Bansal, M., 'G-quadruplex structure can be stable with only some coordination sites being occupied by cations: a six-nanosecond molecular dynamics study'. *J. Phys. Chem. B*, **2001**. 105(31): pp. 7572-8.
150. Srinivasan, J., Cheatham, T.E., III, Ciepak, P., Kollman, P.A., and Case, D., 'Continuum solvent studies of the stability of DNA, RNA and phosphoramidate-DNA helices'. *J. Am. Chem. Soc.*, **1998**. 120(37): pp. 9401-9.
151. Cheatham, T.E., III, Miller, J., Kollman, P.A., and Case, D.A., 'Continuum solvent studies of the stability of RNA hairpin loops and helices'. *J. Biomol. Struct. Dyn.*, **1998**. 16: pp. 671-682.
152. Lee, M.R., Duan, Y., and Kollman, P.A., 'Use of MM-PB/SA in estimating the free energies of proteins: applications to native, intermediates, and unfolded villin headpiece'. *Prot: Struct. Funct. and Gen.*, **2000**. 39: pp. 309.
153. Fadrna, E., Spackova, N., Stefl, R., Koca, J., Cheatham, T.E., III, and Sponer, J., 'Molecular dynamics simulations of guanine quadruplex loops: advances and force field limitations'. *Biophysical Journal*, **2004**. 87: pp. 227-242.
154. Atkins, P. and de Paula, J., *Physical Chemistry*. 7 ed. 2004, Oxford: Oxford University Press. 1180.
155. Jing, N., Rando, R.F., Pommier, Y., and Hogan, M.E., 'Ion selective folding of loop domains in a potent anti-HIV oligonucleotide'. *Biochemistry*, **1997**. 36(41): pp. 12498-505.
156. Pinnavaia, T.J., Marshall, C.L., Mettler, C.M., Fisk, C.L., Miles, H.T., and Becker, E.D., 'Alkali metal ion specificity in the solution ordering of a nucleotide, 5' guanosine monophosphate'. *J. Am. Chem. Soc.*, **1978**. 100(11): pp. 3625-7.
157. Lakowicz, J.R., *Principles of Fluorescence*. 2nd ed. 1999, New York, NY: Plenum Publishing Corp. 698.
158. The Fly Consortium, 'The flybase database of the Drosophila genome projects and community literature http:/flybase.org'. *Nucl. Ac. Res.*, **2003**. 31: pp. 172-175.
159. Kitamoto, T., Wang, W., and Salvaterra, P.M., 'Structure and organisation of the Drosophila cholinergic locus'. *J. Biol. Chem.*, **1998**. 273(5): pp. 2706-2713.
160. Salvaterra, P.M. and McCaman, R.E., 'Choline acetyltransferase and acetylcholine levels in *Drosophila melanogaster*: a study using teo temperature-sensitive mutants'. *J. Neurosci*, **1985**. 5(4): pp. 903-910.
161. Lipinski, C.A., Lombardo, F., Dominy, B.W., and Feeney, P.J., 'Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings'. *Adv. Drug Delivery Rev.*, **1997**. 23: pp. 4-25.

162. Breslauer, K.J., Frank, R., Blocker, H., and Marky, L.A., 'Predicting DNA duplex stability from the base sequence'. *Proc. Natl. Acad. Sci. USA*, **1986**. 83: pp. 3746-3750.
163. Staden, R., 'Methods of calculating the probabilities of finding patterns in sequences'. *Comput. Applic. Biosci.*, **1989**. 5: pp. 89-96.
164. Sewell, R.F. and Durbin, R., 'Method of calculation of probability of matching a bounded regular expression in a random data string'. *J. Comp. Biol.*, **1995**. 2(1): pp. 25-31.
165. Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G., *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. 1998, Cambridge, UK: Cambridge University Press. 368.
166. Elton, R.A., 'Doublet frequencies in sequenced nucleic acids'. *J. Mol. Evol.*, **1975**. 4(4): pp. 323-46.
167. Bird, A.P., 'DNA methylation and the frequency of CpG in animal DNA'. *Nucl. Ac. Res.*, **1980**. 8(7): pp. 1499-1504.
168. Pan, B., Xiong, Y., Shi, K., and Sundaralingam, M., 'Crystal Structure of a Bulged RNA Tetraplex at 1.1 Å Resolution: Implications for a Novel Binding Site in RNA Tetraplex'. *Structure*, **2003**. 11: pp. 1423-30.
169. Agresti, A., *Categorical Data Analysis*. 2nd Ed. ed. 2002, Hoboken, NY: Wiley.
170. Park, G.H., Plummer, H.K.I., and Krystal, G.W., 'Selective Sp1 binding is critical for maximal activity of the human *c-kit* promoter'. *Blood*, **1998**. 92(11): pp. 4138-9.
171. Grant, S.F.A., Reid, D.M., Blake, G., Herd, R., Fogelman, I., and Ralston, S.H., 'Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I alpha 1 gene'. *Nat Genet.*, **1996**. 14(2): pp. 203-205.
172. Mann, V., Hobson, E.E., Li, B., Stewart, T.L., Grant, S.F.A., Robins, S.P., Aspden, R.M., and Ralston, S.H., 'A *COL1A1* Sp1 binding site polymorphism predisposes to osteoporitic fracture by affecting bone density and quality'. *J. Clin. Invest.*, **2001**. 107(7): pp. 899-907.
173. Lania, L., Majello, B., and de Luca, P., 'Transcriptional regulation by the Sp family proteins'. *Int. J. Biochem. Cell Biol.*, **1997**. 29(12): pp. 1313-1323.
174. Suske, G., 'The Sp-family of transcription factors'. *Genes Dev.*, **1999**. 238: pp. 291-300.
175. Berg, J.M., 'Sp1 and the subfamily of zinc finger proteins with guanine-rich binding sites'. *Proc. Natl. Acad. Sci. USA*, **1992**. 89: pp. 11109-11110.
176. Wingender, E., Chen, X., Fricke, E., Geffers, R., Hehl, R., Liebich, I., Krull, M., Maytys, V., Michael, H., Ohnäuser, R., Prüß, M., Schacherer, F., Thiele, S., and Urbach, S., 'The TRANSFAC system on gene expression regulation'. *Nucl. Ac. Res.*, **2001**. 29: pp. 281-3.
177. Hapgood, J.P. and Patterton, D., 'Purification of an oligo(dG).oligo(dC)-binding sea urchin nuclear protein, suGF1: a family of G-string factors involved in gene regulation during development'. *Mol. Cell Biol.*, **1994**. 14(2): pp. 1402-9.
178. Arthanari, H. and Bolton, P.H., 'Functional and dysfunctional roles of quadruplex DNA in cells'. *Chem. & Biol.*, **2001**. 8: pp. 221-230.